Counterfactual Analysis and Target Setting in Benchmarking

Bogetoft, Peter; Ramírez-Ayerbe, Jasone; Romero Morales, Dolores

Document Version
Submitted manuscript

Published in:
European Journal of Operational Research

DOI:
10.1016/j.ejor.2024.01.005

Publication date:
2024

License
CC BY-NC-ND

Citation for published version (APA):
Bogetoft, P., Ramírez-Ayerbe, J., & Romero Morales, D. (2024). Counterfactual Analysis and Target Setting in Benchmarking. *European Journal of Operational Research*, *315*(3), 1083-1095. https://doi.org/10.1016/j.ejor.2024.01.005

Journal Pre-proof

Counterfactual analysis and target setting in benchmarking

Peter Bogetoft, Jasone Ramírez-Ayerbe, Dolores Romero Morales

PII: S0377-2217(24)00006-7
DOI: https://doi.org/10.1016/j.ejor.2024.01.005
Reference: EOR 18834

To appear in: *European Journal of Operational Research*

Received date : 4 July 2023
Accepted date : 4 January 2024

Please cite this article as: P. Bogetoft, J. Ramírez-Ayerbe and D.R. Morales, Counterfactual analysis and target setting in benchmarking. *European Journal of Operational Research* (2024), doi: https://doi.org/10.1016/j.ejor.2024.01.005.

Counterfactual Analysis and Target Setting in Benchmarking

Peter Bogetoft[a], Jasone Ramírez-Ayerbe[b,*], Dolores Romero Morales[a]

[a]*Department of Economics, Copenhagen Business School, Frederiksberg, Denmark*
[b]*Instituto de Matemáticas de la Universidad de Sevilla, Seville, Spain*

Abstract

Data Envelopment Analysis (DEA) allows us to capture the complex relationship between multiple inputs and outputs in firms and organizations. Unfortunately, managers may find it hard to understand a DEA model and this may lead to mistrust in the analyses and to difficulties in deriving actionable information from the model. In this paper, we propose to use the ideas of target setting in DEA and of counterfactual analysis in Machine Learning to overcome these problems. We define DEA counterfactuals or targets as alternative combinations of inputs and outputs that are close to the original inputs and outputs of the firm and lead to desired improvements in its performance. We formulate the problem of finding counterfactuals as a bilevel optimization model. For a rich class of cost functions, reflecting the effort an inefficient firm will need to spend to change to its counterfactual, finding counterfactual explanations boils down to solving Mixed Integer Convex Quadratic Problems with linear constraints. We illustrate our approach using both a small numerical example and a real-world dataset on banking branches.

Keywords: Data Envelopment Analysis, Benchmarking, DEA Targets, Counterfactual Explanations, Bilevel Optimization

1. Introduction

In surveys among business managers, benchmarking is consistently ranked as one of the most popular management tools [38, 39]. The core of benchmarking is relative performance evaluation. The performance of one entity is compared to that of a group of other entities. The evaluated "entity" can be a firm, organization, manager, product or process. In the following, it will be referred to simply as a Decision Making Unit (DMU).

There are many benchmarking approaches and they can serve different purposes, such as, facilitating *learning*, *decision making* and *incentive design*. Some approaches are very simple and rely on the comparison of a DMU's Key Performance Indicators (KPIs) to those of a selected peer group of DMUs. These KPIs are basically partial productivity measures

*Corresponding author
Email addresses: pb.eco@cbs.dk (Peter Bogetoft), mrayerbe@us.es (Jasone Ramírez-Ayerbe), drm.eco@cbs.dk (Dolores Romero Morales)

(e.g., labour productivity, yield per hectare, etc.). This makes KPI based benchmarking easy to understand, but also potentially misleading by ignoring the role of other inputs and outputs in real DMUs. More advanced benchmarking approaches rely on frontier models using mathematical programming, e.g., Data Envelopment Analysis (DEA), and Econometrics, e.g., Stochastic Frontier Analysis (SFA), and they allow us to explicitly model the complex interaction between the multiple inputs and outputs among best-practice DMUs, cf. e.g. [9, 12, 36, 46].

In this paper we focus on DEA based benchmarking. To construct the best practice performance frontier and evaluate the efficiency of a DMU relative to this frontier, DEA introduces a minimum of production economic regularities, typically convexity, and uses linear or mixed integer programming to capture the relationship between multiple inputs and outputs of a DMU. In this sense, and in the eyes of the modeller, the method is well-defined and several of the properties of the model will be understandable from the production economic regularities. Still, from the point of view of the evaluated DMUs, the model will appear very much like a black box. Understanding a multiple input and multiple output structure is basically difficult. Also, in DEA, there is no explicit formula showing the impact of specific inputs on specific outputs as in SFA or other econometrics based approaches. This has led some researchers to look for extra information and structure of DEA models, most notably by viewing the black box as a network of more specific process, cf. e.g. [15, 19, 30].

The black box nature of DEA models may lead to some algorithm aversion and mistrust in the model, and to difficulties in deriving actionable information from the model beyond the efficiency scores. To overcome this and to get insights into the functioning of a DEA model, there are several strands of literature and tools that can be useful. The Multiple Criteria Decision Making (MCDM) literature has developed several ways in which complicated sets of alternatives can be explored and presented to a decision maker. Also, in DEA, there is already a considerable literature on finding targets that a firm can investigate in attempts to find attractive alternative production plans. Last, but not least, it may be interesting to look for counterfactual explanations much like they are used in machine learning.

In this paper, we propose the use of counterfactual and target analyses to understand and explain the efficiencies of individual DMUs, to learn about the estimated best practice technology, and to help answer what-if questions that are relevant in operational, tactical and strategic planning efforts [7]. In a DEA context, counterfactual and target analyses can help with learning, decision making and incentive design. In terms of learning, the DMU may be interested to know what simple changes in features (inputs and outputs) lead to a higher efficiency level. In the application investigated in our numerical section, this can be, for instance, how many credit officers or tellers a bank branch should remove to become fully efficient. This may help the evaluated DMU learn about and gain trust in the underlying modelling. In terms of decision making, targets and counterfactual explanations may help

guide the decision process by offering the smallest, the most plausible and actionable, and the least costly changes that lead to a desired boost in performance. It depends on the context how to define the least costly, or the most plausible or actionable improvement paths. In some cases it may be easier to reduce all inputs more or less the same (lawn mowing), while in other cases certain inputs should be reduced more aggressively than others, cf. [2]. Referring back to the application in the numerical section, reducing the use of different labor inputs could for example take into account the power of different labor unions and the capacity of management to struggle with multiple employee groups simultaneously. Lastly, targets and counterfactual explanations may be useful in connection with incentive provisions. DEA models are routinely used by regulators of natural monopoly networks to incentivize cost reductions and service improvement, cf. e.g. [29] and later updates in [1, 7]. Regulated firms will naturally look for the easiest way to accommodate the regulator's efficiency thresholds. Counterfactual explanations may in such cases serve to guide the optimal strategic responses to the regulator's requirements.

Unfortunately, it is not an entirely trivial task to properly determine targets and construct counterfactual explanations in a DEA context. We need to find alternative solutions that are in some sense close to the existing input-output combination used by a DMU. This involves finding "close" alternatives in the complement of a convex set [44]. In this paper, we investigate different ways to measure the closeness between a DMU and its counterfactual DMU, or the cost of moving from an existing input-output profile to an alternative target. In particular, we suggest to use combinations of ℓ_0, ℓ_1, and ℓ_2 norms. We also consider both changes in input and output features and show how to formulate the problems in DEA models with different returns to scale assumptions. We show how determining targets and constructing counterfactual explanations leads to a bilevel optimization model, that can be reformulated as a Mixed Integer Convex Quadratic Problem with linear constraints. We illustrate our approach on both a small numerical example as well as a large scale real-world dataset involving bank branches.

The outline of the paper is as follows. In Section 2 we review the relevant literature. In Section 3 we introduce the necessary DEA notation for constructing targets and counterfactual explanations, as well as a small numerical example. In Section 4 we describe our bilevel optimization formulation and its reformulation as a Mixed Integer Convex Quadratic Problem with linear constraints. In Section 5 we illustrate our approach with real-world data on bank branches. We end the paper with conclusions in Section 6. In the Appendix, we extend the analysis by investigating alternative returns to scale and by investigating changes in the outputs rather than the inputs.

2. Background and Literature

In this section, we give some background on DEA benchmarking, in particular on directional and interactive benchmarking, on target setting in DEA, and on counterfactual

analysis from interpretable machine learning.

Data Envelopment Analysis, DEA, was first introduced in [13, 14] as a tool for measuring efficiency and productivity of decision making units, DMUs. The idea of DEA is to model the production possibilities of the DMUs and to measure the performance of the individual DMUs relative to the production possibility frontier. The modelling is based on observed practices that form activities in a Linear Programming (LP) based activity analysis model.

Most studies use DEA models primarily to measure the relative efficiency of the DMUs. The benchmarking framework, the LP based activity analysis model, does however allow us to explore a series of other questions. In fact, the benchmarking framework can serve as a learning lab and decision support tool for managers. In the DEA literature, this perspective has been emphasized by the idea of interactive benchmarking. Interactive benchmarking and associated easy to use software has been used in a series of applications and consultancy projects, cf. e.g. [7]. The idea is that a DMU can search for alternative and attractive production possibilities and hereby learn about the technology, explore possible changes and trade-offs and look for least cost changes that allow for necessary performance improvements, cf. also our discussion of learning, decision making and incentive and regulation applications in the introduction.

One way to illustrate the idea of interactive benchmarking is as in Figure 1 below. A DMU has used two inputs to produce two outputs. Based on the data from other DMUs, an estimate of the best practice technology has been established as illustrated by the piecewise linear input and output isoquants. The DMU may now be interested in exploring alternative paths towards best practices. One possibility is to save a lot of input 2 and somewhat less of input 1, i.e., to move in the direction $\boldsymbol{d}_x$ illustrated by the arrow in the left panel. If the aim is to become fully efficient, this approach suggests that the DMU instead of the present (input,output) combination $(\boldsymbol{x}, \boldsymbol{y})$ should consider the alternative $(\hat{\boldsymbol{x}}, \boldsymbol{y})$. A similar logic could be used on the output side keeping the inputs fixed as illustrated in the right panel where we assume that more of a proportional increase in the two outputs is strived at. Of course, in reality, one can combine also changes in the inputs and outputs.

Formally, the directional distance function approach, sometimes referred to as the excess problem, requires solving the following mathematical programming problem

$$\max \quad \{e \mid (\boldsymbol{x} - e\boldsymbol{d}_x, \boldsymbol{y} + e\boldsymbol{d}_y) \in T^*\}, \tag{DIR}$$

where $\boldsymbol{x}$ and $\boldsymbol{y}$ are the present values of the inputs and output vectors, $\boldsymbol{d}_x$ and $\boldsymbol{d}_y$ are the improvement directions in input and output space, T^* is the estimated set of feasible (input,output) combinations, and e is the magnitude of the movement.

In the DEA literature, the direction $(\boldsymbol{d}_x, \boldsymbol{d}_y)$ is often thought as parameters that are given and the excess as one of many possible ways to measure distance to the frontier. A few authors have advocated that some directions are more natural than others and there have been attempts to endogenize the choice of this direction, cf. e.g. [8, 20, 21, 37, 47].

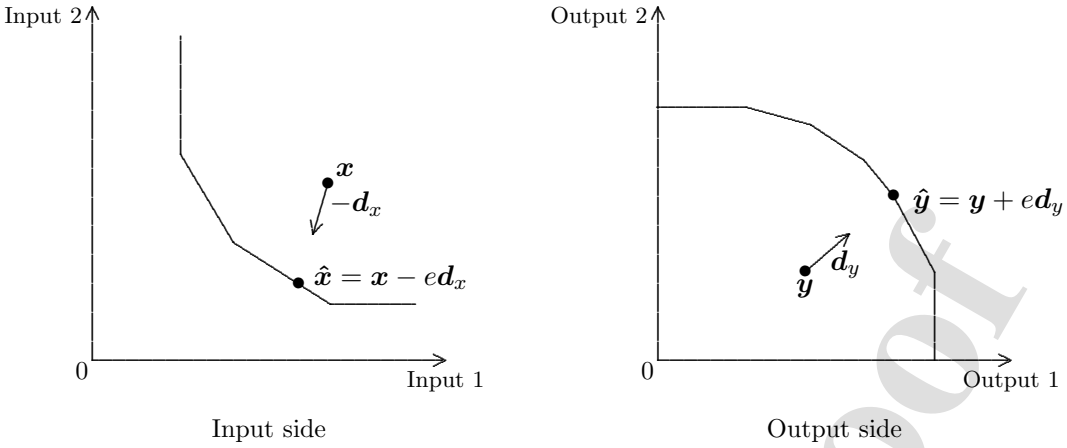


Figure 1: Directional search for an alternative production plan to $(\boldsymbol{x}, \boldsymbol{y})$ along $(\boldsymbol{d}_x, \boldsymbol{d}_y)$ using (DIR)

One can also think of the improvement directions as reflecting the underlying strategy of the DMU or simply as a steering tool that the DMU uses to create one or more interesting points on the frontier.

Figure 2 illustrates the real-world example involving bank branches from the application section. The analysis is here done using the directional distance function approach (DIR) as implemented in the so-called Interactive Benchmarking software, cf. [7]. The search "Direction" is chosen by adjusting the horizontal handles for each input and output and is expressed in percentages of the existing inputs and outputs. The resulting best practice alternative is illustrated in the "Benchmark" column. We see that the DMU in this example expresses an interest in reducing Supervision and Credit personnel but simultaneously seeks to increase the number of personal loan accounts.

Variable	Direction		Present Value	Benchmark	Performance	Type
Teller personnel		0	4.62	4.62	100%	I
Typing personnel		0	0.95	0.95	100%	I
Accountng and Ledgers personnel		0	2.90	2.90	100%	I
Supervision personnel		65	1.58	0.87	55%	I
Credit personnel		25	3.03	2.51	83%	I
Term accounts		0	2487.00	2487.00	100%	O
Pers loan accounts		49	97.00	129.88	75%	O
Commercial loan accounts		0	640.00	640.00	100%	O

Figure 2: Directional search in Interactive Benchmarking software. Real-world dataset of bank branches in Section 5

Applications of interactive benchmarking have typically been in settings where the DMU in a trial-and-error like process seeks alternative production plans. Such processes can certainly be useful in attempts to learn about and gain trust in the modelling, to guide decision making and to find the perfomance enhancing changes that a DMU may find relatively easy to implement. From the point of view of Multiple Criteria Decision Making

(MCDM) we can think of such processes as based on progressive articulation of preferences and alternatives, cf. e.g. the taxonomy of MCDM methods suggested in [40].

It is clear from this small example, however, that the use of an interactive process guided solely by the DMU may not always be the best approach. If there are more than a few inputs and outputs, the process can become difficult to steer towards some underlying optimal compromise between the many possible changes in inputs and outputs. In such cases, the so-called prior articulation of preferences may be more useful. If the DMU can express its preferences for different changes, e.g., as a cost of change function $C((\boldsymbol{x},\boldsymbol{y}),(\boldsymbol{x}^*,\boldsymbol{y}^*))$ giving the cost of moving from the present production plan $(\boldsymbol{x},\boldsymbol{y})$ to any new production plan $(\boldsymbol{x}^*,\boldsymbol{y}^*)$, then a systematic search for the optimal change is possible. The approach of this paper is based on this idea. We consider a class of cost functions and show how to find optimal changes in inputs and outputs using bilevel optimization. In this sense, it corresponds to endogenizing the directional choice so as to make the necessary changes in inputs and outputs as small as possible. Of course, by varying the parameters of the cost function, one can also generate a reasonably representative set of alternative production plans that the DMU can then choose from. This would correspond to the idea of a prior articulation of alternatives approach in the MCDM taxonomy.

The idea of introducing a cost of change function to guide the search for alternative production plans is closely related to the idea of targets in DEA. At a general level, a target is here understood as an alternative production plan that a DMU should move to.[1] There has been a series of interesting DEA papers on the determination of targets using the principle of least action, see for example [4] and the references in here. In [4], the authors explicitly introduce the principle of least action referring to the idea in physics that nature always finds the most efficient course of action. The general argument underlying these approaches is that an inefficient firm should achieve technical efficiency with a minimum amount of effort. Different solutions have been proposed using different distance measures or what we call the cost of change. In many papers, this corresponds to minimizing the distance to the efficient frontier in contrast to the traditional efficiency measurement problem, where we are looking for the largest possible savings or the largest possible expansions of the services provided. A good example is [5]. In this sense, our idea of finding close counterfactuals fits nicely into the DEA literature.

The choice of targets has also been discussed in connection with the slack problem in radial DEA measures. A Farrell projection may not lead to a Pareto efficient point and in a second stage, it is therefore common to discuss close alternatives that are fully Pareto

[1] In [6], the authors distinguish between setting targets and benchmarking in the sense that targets are the coordinates of a projection point, which is not necessarily an observed DMU, whereas benchmarks are real observed DMUs. This distinction can certainly be relevant is several contexts, but it is not how we use targets here. We use benchmarking as the general term for relative performance comparison, and target to designate an alternative production plan that a DMU should choose to improve performance in the easiest possible way.

efficient. Again, different solutions – using for example constraints or the determination of all full facets – have been proposed, cf. [3]. Identifying all facets and measuring distance to these like in [43] is theoretically attractive but computationally cumbersome in most applications.

Our approach can be seen as a generalization of the literature on targets. In particular, if $E^* = 1$ and the considered cost function coincides with a norm or with a typical technical efficiency measure, then the previous DEA target approaches are particular cases of the general approach introduced in this paper. We formulate the target setting problem for general cost-of-change functions using a bilevel program[2], and we reformulate the constraints to get tractable mathematical optimization problems. Using combinations of ℓ_0, ℓ_1 and ℓ_2 norms, as we do in the illustrations, the resulting problems are Mixed Integer Convex Quadratic Problems with linear constraints. It is worthwhile to note also, that we do not necessarily require the target to be Pareto efficient, allowing for the possibility that a DMU may not seek to become fully efficient but, for example, just 90% Farrell efficient which also implies that targets may be on non-full facets.

In interpretable machine learning [16, 41], counterfactual analysis is used to explain the predictions made for individual instances [25, 31, 45]. Machine learning approaches like Deep Learning, Random Forests, Support Vector Machines, and XGBoost are often seen as powerful tools in terms of learning accuracy but also as black boxes in terms of how the model arrives at its outcome. Therefore, regulations from, among others the EU, are enforcing more transparency in the so-called field of algorithmic decision making [18, 24]. There is a paramount of tools being developed in the nascent field of explainable artificial intelligence to help understand how tools in machine learning and artificial intelligence make decisions [32, 33, 34]. The focus of this paper is on counterfactual analysis tools. The starting point is an individual instance for which the model predicts an undesired outcome. In counterfactual analysis, one is interested in building an alternative instance, the so-called counterfactual instance, revealing how to change the features of the current instance so that the model predicts a desired outcome for the counterfactual instance. The counterfactual explanation problem is written as a mathematical optimization problem. To define the problem, one needs to model the feasible space, a cost function measuring the cost of the movement from the current instance to the counterfactual one, and a set of constraints that ensures that the counterfactual explanation is predicted with the desired outcome. In general, the counterfactual explanation problem reads as a constrained nonlinear problem but, for score-based classifiers and cost functions defined by a convex combination of the norms ℓ_0, ℓ_1 and ℓ_2, equivalent Mixed Integer Linear Programming or Mixed Integer Convex Quadratic with Linear Constraints formulations can be defined, see, e.g., [11, 22, 35].

[2]The idea of using a bilevel linear programming approach has also appeared in the DEA literature. It should be noted in particular that [3] proposed to resort to a bilevel linear programming model when strictly efficient targets are to be identified using the Russel output measure to capture cost-of-change.

In the following, we combine the ideas of DEA, least action targets, and counterfactual explanations. We formulate and solve bilevel optimization models to determine "close" alternative production plans or counterfactual explanations in DEA models that lead to desired relative performance levels and also take into account the strategic preferences of the entity.

3. The Setting

We consider $K+1$ DMUs (indexed by k), using I inputs, $\boldsymbol{x}^k = (x_1^k, \ldots, x_I^k)^\top \in \mathbb{R}_+^I$, to produce O outputs, $\boldsymbol{y}^k = (y_1^k, \ldots, y_O^k)^\top \in \mathbb{R}_+^O$. Hereafter, we will write $(\boldsymbol{x}^k, \boldsymbol{y}^k)$ to refer to production plan of DMU k, $k = 0, 1, \ldots, K$.

Let T be the technology set, with

$$T = \{(\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{R}_+^I \times \mathbb{R}_+^O \quad | \quad \boldsymbol{x} \text{ can produce } \boldsymbol{y}\}.$$

We will initially estimate T by the classical DEA model. It determines the empirical reference technology T^* as the smallest subset of $\mathbb{R}_+^I \times \mathbb{R}_+^O$ that contains the actual $K+1$ observations, and satisfies the classical DEA regularities of convexity, free-disposability in inputs and outputs, and Constant Returns to Scale (CRS). It is easy to see that the estimated technology can be described as:

$$T^*(\text{CRS}) = \{(\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{R}_+^I \times \mathbb{R}_+^O \quad | \quad \exists \boldsymbol{\lambda} \in \mathbb{R}_+^{K+1} : \boldsymbol{x} \geq \sum_{k=0}^{K} \lambda^k \boldsymbol{x}^k, \boldsymbol{y} \leq \sum_{k=0}^{K} \lambda^k \boldsymbol{y}^k\}.$$

To measure the efficiency of a firm, we will initially use the so-called Farrell input-oriented efficiency. It measures the efficiency of a DMU, say DMU 0, as the largest proportional reduction E^0 of all its inputs $\boldsymbol{x}^0$ that allows the production of its present outputs $\boldsymbol{y}^0$ in the technology T^*. Hence, it is equal to the optimal solution value of the following LP formulation

$$\begin{aligned} \min_{E, \lambda^0, \ldots, \lambda^K} \quad & E && \text{(DEA)} \\ \text{s.t.} \quad & E\boldsymbol{x}^0 \geq \sum_{k=0}^{K} \lambda^k \boldsymbol{x}^k \\ & \boldsymbol{y}^0 \leq \sum_{k=0}^{K} \lambda^k \boldsymbol{y}^k \\ & 0 \leq E \leq 1 \\ & \boldsymbol{\lambda} \in \mathbb{R}_+^{K+1}. \end{aligned}$$

This DEA model has $K+2$ decision variables, I linear input constraints and O linear output constraints. Hereafter, we will refer to the optimal objective value of (DEA), say E^0, as the

efficiency of DMU 0.

In the following, and assuming that firm 0 with production plan $(\boldsymbol{x}^0, \boldsymbol{y}^0)$ is not fully efficient, $E^0 < 1$, we will show how to calculate a *counterfactual explanation* with a desired efficiency level $E^* > E^0$, i.e., the minimum changes needed in the inputs of the firm, $\boldsymbol{x}^0$, in order to obtain an efficiency E^*. Given a cost function $C(\boldsymbol{x}^0, \hat{\boldsymbol{x}})$ that measures the cost of moving from the present inputs $\boldsymbol{x}^0$ to the new counterfactual inputs $\hat{\boldsymbol{x}}$, and a set $\mathcal{X}(\boldsymbol{x}^0)$ defining the feasible space for $\hat{\boldsymbol{x}}$, the counterfactual explanation for $\boldsymbol{x}^0$ is found solving the following optimization problem:

$$\begin{aligned} \min_{\hat{\boldsymbol{x}}} \quad & C(\boldsymbol{x}^0, \hat{\boldsymbol{x}}) \\ \text{s.t.} \quad & \hat{\boldsymbol{x}} \in \mathcal{X}(\boldsymbol{x}^0) \\ & (\hat{\boldsymbol{x}}, \boldsymbol{y}^0) \quad \text{has at least an efficiency of } E^*. \end{aligned}$$

With respect to $C(\boldsymbol{x}^0, \hat{\boldsymbol{x}})$, different norms can be used to measure the difficulty of changing the inputs. A DMU may, for example, be interested to minimize the sum of the squared deviations between the present and the counterfactual inputs. We model this using the squared Euclidean norm ℓ_2^2. Likewise, there may be an interest in minimizing the absolute value of the deviations, which we can proxy using the ℓ_1 norm, or the number of inputs changed, which we can capture with the ℓ_0 norm. When it comes to $\mathcal{X}(\boldsymbol{x}^0)$, this would include the nonnegativity of $\hat{\boldsymbol{x}}$, as well as domain knowledge specific constraints. With this approach, we detect the most important inputs in terms of the impact they have on the DMU's efficiency, and with enough flexibility to consider different costs of changing depending on the DMU's characteristics.

In the next section, we will show that finding counterfactual explanations involves solving a bilevel optimization problem of minimizing the changes in inputs and solving the above DEA problem at the same time. In the Appendix, we will also discuss how the counterfactual analysis approach can be extended to other technologies and to other efficiency measures like the output-oriented Farrell efficiency and other DEA technologies.

Before turning to the details of the bilevel optimization problem, it is useful to illustrate the idea of counterfactual explanations using a small numerical example. Suppose we have four firms with the inputs, outputs, and Farrell input efficiencies as in Table 1. The efficiency has been calculated solving the classical DEA model with CRS, namely (DEA). In this example, firms 1 and 2 are fully efficient, whereas firms 3 and 4 are not.

Firm	x_1	x_2	y	E
1	0.50	1	1	1
2	1.50	0.50	1	1
3	1.75	1.25	1	0.59
4	2.50	1.25	1	0.50

Table 1: Inputs, outputs and corresponding Farrell input-efficiency of 4 different firms

First, we want to know the changes needed in $\boldsymbol{x}^3$ for firm 3 to have a new efficiency E^* of at least 80%. Since we only have two inputs, we can illustrate this graphically as in Figure 3a. The results are shown in Table 2 for different cost functions. It can be seen that we in all cases get exactly 80% efficiency with the new inputs. We see from column ℓ_2^2 that the Farrell solution is further away from the original inputs than the counterfactual solution based on the Euclidean norm. To the extent that difficulties of change is captured by the ℓ_2^2 norm, we can conclude that the Farrell solution is not ideal. Moreover, in the Farrell solution one must by definition change both inputs, see column ℓ_0. Using a cost function combining the ℓ_0 norm and the squared Euclidean norm, denoted by $\ell_0 + \ell_2$, one penalizes the number of inputs changed. With this we detect the one input that should change in order to obtain a higher efficiency, namely the second input. In contexts like negotiations with various input suppliers, it is often more practical to focus negotiations on just one or a select few inputs, instead of dealing with all inputs at the same time.

Cost function	$\hat{x}_1$	$\hat{x}_2$	y	E	ℓ_2^2	ℓ_0
Farrell	1.29	0.92	1	0.8	0.32	2
$\ell_0 + \ell_2$	1.75	0.69	1	0.8	0.31	1
ℓ_2	1.53	0.80	1	0.8	0.25	2

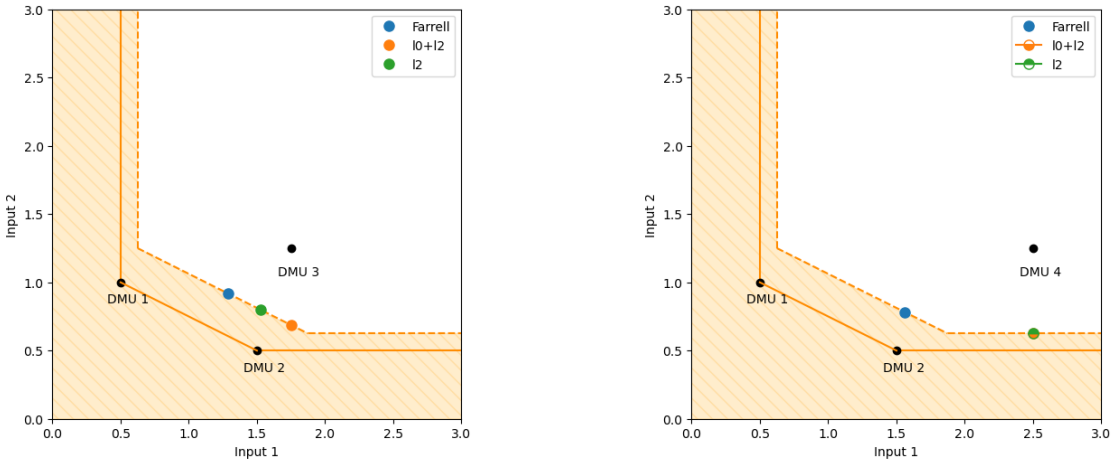
Table 2: Counterfactual explanations for firm 3 in Table 1 imposing $E^* = 0.8$ and different cost functions

Let us now focus on firm 4 and again find a counterfactual instance with at least 80% efficiency. The results are shown in Table 3 and Figure 3b. Notice how in the Farrell case one obtains again the farthest solution and also the least sparse from the three of them. As for the counterfactual explanations with our methodology, the inputs nearest to the original DMU that give us the desired efficiency are in a non full-facet of the efficiency frontier. By using the Farrell to measure the desired efficiency level, we only need to change one input, namely, the second input, and can have "slack" in the first input. We here deviate from [3], in which the authors look for targets on the strongly efficient frontier, i.e., without slack.

Cost function	$\hat{x}_1$	$\hat{x}_2$	y	E	ℓ_2^2	ℓ_0
Farrell	1.56	0.78	1	0.8	1.10	2
$\ell_0 + \ell_2$	2.50	0.63	1	0.8	0.39	1
ℓ_2	2.50	0.63	1	0.8	0.39	1

Table 3: Counterfactual explanations for firm 4 in Table 1 imposing $E^* = 0.8$ and different cost functions

In Figure 3, the space where we search for the counterfactual explanation is shaded. Although in these illustrations the frontier is explicitly given, in general, the frontier points are convex combinations of the observed DMUs, and obtaining the isoquants is not easy. Therefore, when finding counterfactual explanations a bilevel optimization is needed to search for "close" inputs in the complement of a convex set.



(a) Explanations for firm 3

(b) Explanations for firm 4

Figure 3: Counterfactual explanations for firms 3 and 4 in Tables 2 and 3 respectively imposing $E^* = 0.8$ and different cost functions

4. Bilevel optimization for counterfactual analysis in DEA

Suppose DMU 0 is not fully efficient, i.e., the optimal objective value of Problem (DEA) is $E^0 < 1$. In this section, we formulate the counterfactual explanation problem in DEA, i.e., the problem that calculates the minimum cost changes in the inputs $\boldsymbol{x}^0$ that make DMU 0 have a higher efficiency. Let $\hat{\boldsymbol{x}}$ be the new inputs of DMU 0 that would make it at least E^* efficient, with $E^* > E^0$. With this, we have defined the counterfactual instance as the one obtained changing the inputs, but in the same sense, we could define it by changing the outputs. This alternative output-based problem will be studied in the Appendix.

Since the values of the inputs are to be changed, the efficiency of the new production plan $(\hat{\boldsymbol{x}}, \boldsymbol{y}^0)$ has to be calculated using Problem (DEA). The counterfactual explanation problem in DEA reads as follows:

$$\min_{\hat{\boldsymbol{x}}} \quad C(\boldsymbol{x}^0, \hat{\boldsymbol{x}}) \tag{1}$$

$$\text{s.t.} \quad \hat{\boldsymbol{x}} \in \mathbb{R}^I_+ \tag{2}$$

$$E \geq E^* \tag{3}$$

$$E \in \underset{\bar{E},\lambda^0,\ldots,\lambda^K}{\arg\min} \left\{ \bar{E} \quad : \quad \bar{E}\hat{\boldsymbol{x}} \geq \sum_{k=0}^{K} \lambda^k \boldsymbol{x}^k, \boldsymbol{y}^0 \leq \sum_{k=0}^{K} \lambda^k \boldsymbol{y}^k, \bar{E} \geq 0, \boldsymbol{\lambda} \in \mathbb{R}^{K+1}_+ \right\}, \tag{4}$$

where in the upper level problem in (1) we minimize the cost of changing the inputs for firm 0, $\boldsymbol{x}^0$, to $\hat{\boldsymbol{x}}$, ensuring nonnegativity of the inputs, as in constraint (2), and that the efficiency is at least E^*, as in constraint (3). The lower level problem in (4) ensures that the efficiency of $(\hat{\boldsymbol{x}}, \boldsymbol{y}^0)$ is correctly calculated. Therefore, as opposed to counterfactual analysis in interpretable machine learning, here we are confronted with a bilevel optimization

problem. Notice also that to calculate the efficiency in the lower level problem in (4), the technology is already fixed, and the new DMU $(\hat{\boldsymbol{x}}, \boldsymbol{y}^0)$ does not take part in its calculation.

In what follows, we reformulate the bilevel optimization problem (1)-(4) as a single-level model, by exploiting the optimality conditions for the lower-level problem. This can be done for convex lower-level problems that satisfy Slater's conditions, e.g., if our lower-level problem was linear. In our case, however, not all the constraints are linear, since in (4) we have the product of decision variables $\bar{E}\hat{\boldsymbol{x}}$. To be able to handle this, we define new decision variables, namely, $F = \frac{1}{\bar{E}}$ and $\beta^k = \frac{\lambda^k}{\bar{E}}$, for $k = 0, \ldots, K$. Thus, (1)-(4) is equivalent to:

$$\begin{aligned} \min_{\hat{\boldsymbol{x}}, F} \quad & C(\boldsymbol{x}^0, \hat{\boldsymbol{x}}) \\ \text{s.t.} \quad & \hat{\boldsymbol{x}} \in \mathbb{R}^I_+ \\ & F \leq F^* \\ & F \in \arg\max_{\bar{F}, \boldsymbol{\beta}} \{\ \bar{F} \quad : \quad \hat{\boldsymbol{x}} \geq \sum_{k=0}^{K} \beta^k \boldsymbol{x}^k, \bar{F}\boldsymbol{y}^0 \leq \sum_{k=0}^{K} \beta^k \boldsymbol{y}^k, \bar{F} \geq 0, \boldsymbol{\beta} \in \mathbb{R}^{K+1}_+ \ \}. \end{aligned} \tag{5}$$

This equivalent bilevel opimization problem can now be reformulated as a single-level model. The new lower-level problem in (5) can be seen as the $\hat{\boldsymbol{x}}$-parametrized problem:

$$\max_{F, \boldsymbol{\beta}} \quad F \tag{6}$$

$$\text{s.t.} \quad \hat{\boldsymbol{x}} \geq \sum_{k=0}^{K} \beta^k \boldsymbol{x}^k \tag{7}$$

$$F\boldsymbol{y}^0 \leq \sum_{k=0}^{K} \beta^k \boldsymbol{y}^k \tag{8}$$

$$F \geq 0 \tag{9}$$

$$\boldsymbol{\beta} \geq \mathbf{0}. \tag{10}$$

The Karush-Kuhn-Tucker (KKT) conditions, which include primal and dual feasibility, stationarity and complementarity conditions, are necessary and sufficient to characterize an optimal solution. Thus, we can replace problem (5) by its KKT conditions. Primal feasibility is given by (7)-(10). Dual feasibility is given by:

$$\boldsymbol{\gamma}_I, \boldsymbol{\gamma}_O, \delta, \boldsymbol{\mu} \geq \mathbf{0}, \tag{11}$$

for the dual variables associated with constraints (7)-(10), where $\boldsymbol{\gamma}_I \in \mathbb{R}^I_+$, $\boldsymbol{\gamma}_O \in \mathbb{R}^O_+$, $\delta \in \mathbb{R}_+$, $\boldsymbol{\mu} \in \mathbb{R}^{K+1}_+$. The stationarity conditions are as follows:

$$\boldsymbol{\gamma}_O^\top \boldsymbol{y}^0 - \delta = 1 \tag{12}$$

$$\boldsymbol{\gamma}_I^\top \boldsymbol{x}^k - \boldsymbol{\gamma}_O^\top \boldsymbol{y}^k - \mu_k = 0 \quad k = 0, \ldots, K. \tag{13}$$

Lastly, we need the complementarity conditions for all constraints (7)-(10). For constraint (7), we have:

$$\gamma_I^i = 0 \quad \text{or} \quad \hat{x}_i - \sum_{k=0}^{K} \beta^k x_i^k = 0 \quad i = 1, \ldots, I. \tag{14}$$

In order to model this disjunction, we will introduce binary variables $u_i \in \{0, 1\}$, $i = 1, \ldots, I$, and the following constraints using the big-M method:

$$\gamma_I^i \leq M_I u_i, \quad \hat{x}_i - \sum_{k=0}^{K} \beta^k x_i^k \leq M_I(1 - u_i), \quad i = 1, \ldots, I, \tag{15}$$

where M_I is a sufficiently large constant.

The same can be done for the complementarity condition for constraint (8), introducing binary variables $v_o \in \{0, 1\}$, $o = 1, \ldots, O$, big-M constant M_O, and constraints:

$$\gamma_O^o \leq M_O v_o, \quad -F y_o^0 + \sum_{k=0}^{K} \beta^k y_o^k \leq M_O(1 - v_o), \quad o = 1, \ldots, O. \tag{16}$$

The complementarity condition for constraint (10) would be the disjunction $\beta^k = 0$ or $\mu_k = 0$. Using the stationarity condition (13) and again the big-M method with binary variables $w_k \in \{0, 1\}$, $k = 0, \ldots, K$, and big-M constant M_f, one obtains the constraints:

$$\beta^k \leq M_f w_k, \quad \boldsymbol{\gamma}_I^\top \boldsymbol{x}^k - \boldsymbol{\gamma}_O^\top \boldsymbol{y}^k \leq M_f(1 - w_k), \quad k = 0, \ldots, K. \tag{17}$$

Finally, for constraint (9) the complementarity condition yields $F = 0$ or $\delta = 0$. Remember that $F = 1/E$, $0 \leq E \leq 1$, thus F cannot be zero by definition and we must impose $\delta = 0$. Using stationarity condition (12), this yields:

$$\boldsymbol{\gamma}_O^\top \boldsymbol{y}^0 = 1. \tag{18}$$

We now reflect on the meaning of these constraints. Notice that constraints (15) and (16) model the slacks of the inputs and outputs respectively, while constraint (17) models the firms that define the frontier, i.e., the firms with which DMU 0 is to be compared. If binary variable $u_i = 1$, then there is no slack in input i, i.e., $\hat{x}_i = \sum_{k=1}^{K} \beta^k x_i^k$, whereas if $u_i = 0$ that means there is. The same happens with binary variable v_o, namely, it indicates whether there is a slack in output o. On the other hand, when $w_k = 1$, then the equality of the dual constraint will hold $\boldsymbol{\gamma}_I^\top \boldsymbol{x}^k = \boldsymbol{\gamma}_O^\top \boldsymbol{y}^k$, i.e., firm k is fully efficient and it is used to define the efficiency of the counterfactual instance. If $w_k = 0$ then $\beta_k = 0$, and firm k is not being used to define the efficiency of the counterfactual instance. Let us go back to the example in the previous section with four firms with 2 inputs and 1 output and several choices of cost function C of changing the inputs. When $C = \ell_2^2$, we can see that firm 3 is compared against firms 1 and 2, while firm 4 is compared against firm 2 only.

Notice that μ_k is only present in (13), thus it is free. In addition, we know that $\delta = 0$. Therefore, we can transform the stationarity conditions (12) and (13) to

$$\boldsymbol{\gamma}_O^\top \boldsymbol{y}^0 = 1 \tag{19}$$

$$\boldsymbol{\gamma}_I^\top \boldsymbol{x}^k - \boldsymbol{\gamma}_O^\top \boldsymbol{y}^k \geq 0 \quad k = 1, \ldots, K \tag{20}$$

$$\boldsymbol{\gamma}_I, \boldsymbol{\gamma}_O \geq \mathbf{0}, \tag{21}$$

that are exactly the constraints in the dual DEA model for the Farrell output efficiency.

The new reformulation of the counterfactual explanation problem in DEA is as follows:

$$\begin{aligned} \min_{\hat{\boldsymbol{x}}, F, \boldsymbol{\beta}, \boldsymbol{\gamma}_I, \boldsymbol{\gamma}_O, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}} \quad & C(\boldsymbol{x}^0, \hat{\boldsymbol{x}}) \\ \text{s.t.} \quad & F \leq F^* \\ & \hat{\boldsymbol{x}} \in \mathbb{R}_+^I \\ & \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w} \in \{0, 1\} \\ & (7) - (10) \quad \text{primal} \\ & (19) - (21) \quad \text{dual} \\ & (15) - (16) \quad \text{slacks} \\ & (17) \quad \text{frontier.} \end{aligned}$$

So far, we have not been very specific about the objective function $C(\boldsymbol{x}^0, \hat{\boldsymbol{x}})$. Different functional forms can be introduced, and this may require the introduction of further variables to implement these.

In Section 3, we approximated the firm's cost-of-change using combinations of the ℓ_0 norm, the ℓ_1 norm, and the squared ℓ_2 norm. They are widely used in machine learning when close counterfactuals are sought in attempt to understand how to getter a more attractive outcome [10]. The ℓ_0 "norm", which strictly speaking is not a norm in the mathematical sense, counts the number of dimensions that has to be changed. The ℓ_1 norm is the absolute value of the deviations. Lastly, ℓ_2^2 is the Euclidean norm, that squares the deviations.

As a starting point, we therefore propose the following objective function:

$$C(\boldsymbol{x}^0, \hat{\boldsymbol{x}}) = \nu_0 \|\boldsymbol{x}^0 - \hat{\boldsymbol{x}}\|_0 + \nu_1 \|\boldsymbol{x}^0 - \hat{\boldsymbol{x}}\|_1 + \nu_2 \|\boldsymbol{x}^0 - \hat{\boldsymbol{x}}\|_2^2, \tag{22}$$

where $\nu_0, \nu_1, \nu_2 \geq 0$. Taking into account that there may be specific product input prices and output prices or that inputs may have varying degrees of difficulty to be changed, one can consider giving different weights to the deviations in each of the inputs.

In order to have a smooth expression of objective function (22), additional decision variables and constraints have to be added to the counterfactual explanation problem in

DEA. To linearize the ℓ_0 norm, binary decision variables ξ_i are introduced. For input i, $\xi_i = 1$ models $x_i^0 \neq \hat{x}_i$, $i = 1, \ldots, I$. Using the big-M method the following constraints are added to our formulation:

$$-M_{\text{zero}}\xi_i \leq x_i^0 - \hat{x}_i \leq M_{\text{zero}}\xi_i, \quad i = 1, \ldots, I \tag{23}$$

$$\xi_i \in \{0, 1\}, \quad i = 1, \ldots, I, \tag{24}$$

where M_{zero} is a sufficiently large constant.

For the ℓ_1 norm we introduce continuous decision variables $\eta_i \geq 0$, $i = 1, \ldots, I$, to measure the absolute values of the deviations, $\eta_i = \left|x_i^0 - \hat{x}_i\right|$, which is naturally implemented by the following constraints:

$$\eta_i \geq x_i^0 - \hat{x}_i, \quad i = 1, \ldots, I \tag{25}$$

$$-\eta_i \leq x_i^0 - \hat{x}_i, \quad i = 1, \ldots, I \tag{26}$$

$$\eta_i \geq 0, \quad i = 1, \ldots, I. \tag{27}$$

Thus, the counterfactual explanation problem in DEA with cost function C in (22), hereafter (CEDEA), reads as follows:

$$\begin{aligned} \min_{\hat{\boldsymbol{x}}, F, \boldsymbol{\beta}, \gamma_I, \gamma_O, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}, \boldsymbol{\eta}, \boldsymbol{\xi}} \quad & \nu_0 \sum_{i=1}^{I} \xi_i + \nu_1 \sum_{i=1}^{I} \eta_i + \nu_2 \sum_{i=1}^{I} \eta_i^2 \\ \text{s.t.} \quad & F \leq F^* \\ & \hat{\boldsymbol{x}} \in \mathbb{R}_+^I \\ & \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w} \in \{0, 1\} \\ & (7) - (10), (15) - (17), (19) - (21), \\ & (23) - (24), (25) - (27). \end{aligned} \tag{CEDEA}$$

Notice that in Problem (CEDEA) we assumed $\mathcal{X}(\boldsymbol{x}^0) = \mathbb{R}_+^I$ as the feasible space for $\hat{\boldsymbol{x}}$. Other relevant constraints for the counterfactual inputs could easily be added, e.g., bounds or relative bounds on the inputs, or inputs that cannot be changed in the short run, say capital expenses, or that represent environmental conditions beyond the control of the DMU.

In the case where only the ℓ_0 and ℓ_1 norms are considered, i.e., $\nu_2 = 0$, the objective function as well as the constraints are linear, while we have both binary and continuous decision variables. Therefore, Problem (CEDEA) can be solved using an Mixed Integer Linear Programming (MILP) solver. Otherwise, when $\nu_2 \neq 0$, Problem (CEDEA) is a Mixed Integer Convex Quadratic model with linear constraints, which can be solved with standard optimization packages. When all three norms are used, Problem (CEDEA) has $3I + K + 2 + O$ continuous variables and $2I + O + K + 1$ binary decision variables. It has $7I + 3O + 3K + 5$ constraints, plus the non-negativity and binary nature of the variables.

The computational experiments show that this problem can be solved efficiently for our real-world dataset.

We can think of the objective function C in different ways.

One possibility is to see it as an instrument to explore the production possibilities. The use of a combinations of the ℓ_0, ℓ_1 and ℓ_2 norms seems natural here. Possible extensions could involve other ℓ_p norms, $\|\boldsymbol{x}^0 - \hat{\boldsymbol{x}}\|_p := \left(\sum_{i=1}^{I} |x_i^0 - \hat{x}_i|^p\right)^{1/p}$. For all $p \in [1, \infty)$, ℓ_p is convex. This makes the use of ℓ_p norms convenient in generalizations of Problem (CEDEA). Of course, arbitrary ℓ_p norms may lead to more complicated implementations in existing softwares since the objective function may no longer be quadratic.

Closely related to the instrumental view of the objective function is the idea of approximations. At least as a reasonable initial approximation of more complicated functions, many objective functions C can be approximated by the form in (22).

To end, one can link the form of C closer to economic theory. In the economic literature there have been many studies on *factor adjustments costs*. It is commonly believed that firms change their demand for inputs only gradually and with some delay, cf. e.g. [28]. For labor inputs, the factor adjustment costs include disruptions to production occurring when changing employment causes workers' assignments to be rearranged. Laying off or hiring new workers is also costly. There are search costs (advertising, screening, and processing new employees); the cost of training (including disruptions to production as previously trained workers' time is devoted to on-the-job instruction of new workers); severance pay (mandated and otherwise); and the overhead cost of maintaining that part of the personnel function dealing with recruitment and worker outflows. Following again [28], the literature on both labor and capital goods adjustments has overwhelmingly relied on one form of C, namely that of symmetric convex adjustment costs much like we use in (22). Indeed, in the case of only one production factor, the most widely used function form is simply the quadratic one. Hall [27] and several others have tried to estimate the costs of adjusting labor and capital inputs. Using a Cobb-Douglas production function, and absent adjustment costs and absent changes in the ratios of factor prices, an increase in demand or in another determinant of industry equilibrium would cause factor inputs to change in the same proportion as outputs. Adjustment costs are introduced as reductions in outputs and are assumed to depend on the squared growth rates in labor and capital inputs - the larger the percentage change, the larger the adjustment costs. Another economic approach to the cost-of-change modelling is to think of *habits*. In firms - as in the private life - habits are useful. In the performance of many tasks, including complicated ones, it is easiest to go into automatic mode and let a behavior unfold. When an efficiency requirement is introduced, habits may need to change and this is costly. The relevance and strength of habit formation has also been studied empirically using panel data, cf. e.g. [17] and the references herein. Habit formation should ideally be considered in a dynamic framework. To keep it simple, we might consider two periods - the past, where $\boldsymbol{x}^0$ was used and the present, where $\hat{\boldsymbol{x}}$ is consumed. The

utility in period two will then typically depend on the difference or ratio of present to past consumption, $\hat{\boldsymbol{x}} - \boldsymbol{x}^0$ or, in the unidimensional case, $\hat{\boldsymbol{x}}/\boldsymbol{x}^0$. Examples of functional forms one can use are provided in for example [23].

5. A banking application

In this section, we illustrate our methodology using real-world data on bank branches, [42], by constructing a collection of counterfactual explanations for each of the inefficient firms that can help them learn about the DEA benchmarking model and how they can improve their efficiency.

The data is described in more detail in Section 5.1, where we consider a model of bank branch production with $I = 5$ inputs and $O = 3$ outputs, and thus a production possibility set in $\mathbb{R}^8_+$, spanned by $K + 1 = 267$ firms. In Section 5.2, we will focus on changing the inputs, and therefore the counterfactual explanations will be obtained with Problem (CEDEA). We will discuss the results obtained with different cost of change functions C, reflecting the effort an inefficient firm will need to spend to change to its counterfactual instance, and different desired levels of efficiency E^*. The Farrell projection discussed in Section 3 is added for reference. The counterfactual analysis sheds light on the nature of the DEA benchmarking model, which is otherwise hard to comprehend because of the many firms and inputs and outputs involved in the construction of the technology.

All optimization models have been implemented using Python 3.8 and as solver Gurobi 9.0 [26]. We have solved Problem (CEDEA) with $M_I = M_0 = M_f = 1000$ and $M_{\text{zero}} = 1$. The validity of $M_{\text{zero}} = 1$ will be shown below. Our numerical experiments have been conducted on a PC, with an Intel R CoreTM i7-1065G7 CPU @ 1.30GHz 1.50 GHz processor and 16 gigabytes RAM. The operating system is 64 bits.

5.1. The data

The data consist of five staff categories and three different types of outputs in the Ontario branches of a large Canadian bank. The inputs are measured as full-time equivalents (FTEs), and the outputs are the average monthly counts of the different transactions and maintenance activities. Observations with input values equal to 0 are removed, leaving us with an actual dataset with 267 branches. Summary statistics are provided in Table 4.

After calculating all the efficiencies through Problem (DEA), one has that 236 firms of the 267 ones are inefficient. Out of those, 219 firms have an efficiency below 90%, 186 below 80%, 144 below 70%, 89 below 60% and 49 below 50%.

5.2. Counterfactual analysis of bank branches

To examine the inefficient firms, we will determine counterfactual explanations for these. Prior to that, we have divided each input by its maximum value across all firms. We notice that this has no impact on the solution since DEA models are invariant to linear

	Mean	Min	Max	Std. dev.
INPUTS				
Teller	5.83	0.49	39.74	3.80
Typing	1.05	0.03	22.92	1.84
Accounting & ledgers	4.69	0.80	65.93	5.13
Supervision	2.05	0.43	38.29	2.66
Credit	4.40	0.35	55.73	6.19
OUTPUTS				
Term accounts	2788	336	22910	2222
Personal loan accounts	117	0	1192	251
Commercial loan accounts	858	104	8689	784

Table 4: Descriptive statistics of the Canadian bank branches dataset in [42]

transformations of inputs and outputs. Also, this makes valid choosing $M_{\text{zero}} = 1$, since the values of all inputs are upper bounded by 1.

We will use three different cost functions, by changing the values of the parameters ν_0, ν_1, ν_2 in (22), as well as two different values of the desired efficiency of the counterfactual instance, namely $E^* = 1$ and 0.8. In the first implementation of the cost function, which we denote $\ell_0 + (\ell_2)$, we use $\nu_0 = 1$, $\nu_2 = 10^{-3}$ and $\nu_1 = 0$, i.e., we will seek to minimize the ℓ_0 norm and only introduce a little bit of the squared Euclidean norm to ensure a unique solution of Problem (CEDEA). In the second implementation, which we call $\ell_0 + \ell_2$, we take $\nu_0 = 1$, $\nu_1 = 0$ and $\nu_2 = 10^5$, such that the squared Euclidean norm has a higher weight than in cost function $\ell_0 + (\ell_2)$. Finally, we denote by ℓ_2 the cost function that focuses on the minimization of the squared Euclidean norm only, i.e., $\nu_0 = \nu_1 = 0$ and $\nu_2 = 1$. The summary of all the cost functions used can be seen in Table 5. Calculations were also done for the ℓ_1 norm, i.e., $\nu_0 = \nu_2 = 0$ and $\nu_1 = 1$, but as the solutions found were similar to those for cost function $\ell_0 + \ell_2$, for the sake of clarity of presentation, they are omitted. We start the discussion of the counterfactual explanations obtained with $E^* = 1$, as summarized in Figures 4-5 and Tables 6-7. We then move on to a less demanding desired efficiency, namely, $E^* = 0.8$. These results are summarized in Figures 6-7 and Tables 8-9.

Cost function	ν_0	ν_1	ν_2
$\ell_0 + (\ell_2)$	1	0	10^{-3}
$\ell_0 + \ell_2$	1	0	10^5
ℓ_2	0	0	1

Table 5: Value of the parameters ν_0, ν_1 and ν_2 in (22) for the different cost functions used

Let us first visualize the counterfactual explanations for a specific firm. Consider, for instance, firm 238, which has an original efficiency of $E^0 = 0.72$. We can visualize the different counterfactual explanations generated by the different cost functions using a spider chart, see Figure 4. In addition to the counterfactual explanations obtained with Problem (CEDEA), we also illustrate the so-called Farrell projection. In the spider chart, each axis represents an input and the original values of the firm corresponds to the outer circle.

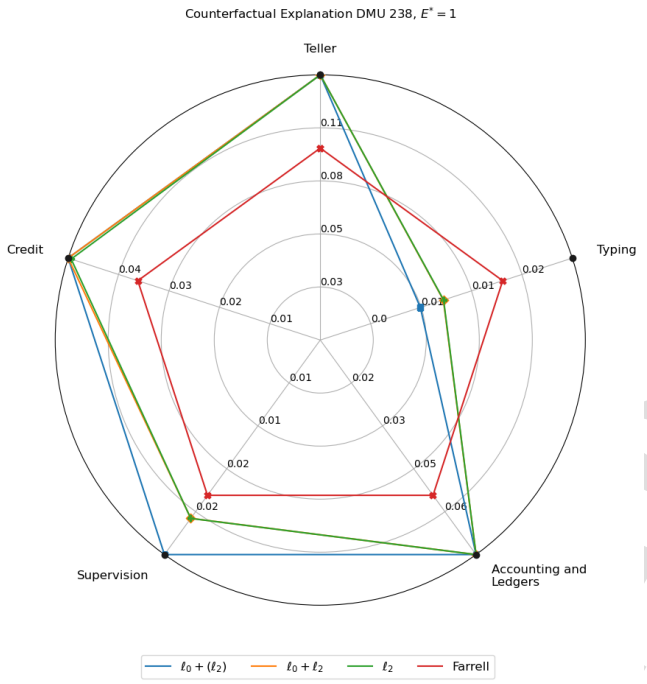


Figure 4: Counterfactual Explanations for firm 238 with Problem (CEDEA) and desired efficiency $E^* = 1$.

Figure 4 shows the different changes needed depending on the cost function used. With the $\ell_0 + (\ell_2)$, where the focus is to mainly penalize the number of inputs changed, we see that only the typing personnel has to be changed, leaving the rest of the inputs unchanged. Nevertheless, because only one input is changed, it has to be decreased by 60% from the original value. The Farrell solution decreases the typing personnel by 28% of its value, but to compensate, it changes the remaining four inputs proportionally. When the $\ell_0 + \ell_2$ cost function is used, the typing personnel keeps on needing to be changed, but the change is smaller, this time by 51% of its value. The supervision personnel needs also to be decreased by 16% of its value, while the rest of the inputs remain untouched. Increasing the weight on the Euclidean norm in the cost function gives us the combination of the two inputs that are crucial to change in order to gain efficiency, as well as the exact amount that they need to be reduced. Finally, using only the Euclidean norm, the typing, supervision and credit personnel are the inputs to be changed, the typing input is reduced slightly less than with the $\ell_0 + \ell_2$ in exchange of reducing just by 1% the credit input. Notice that the teller and accounting and ledgers personnel are never changed in the counterfactual explanations generated by our methodology, which leads us to think that these inputs are not the ones leading firm 238 to have its original low efficiency.

The analysis above is for a single firm. We now present some statistics about the counterfactual explanations obtained for all the inefficient firms. Recall that these are 236 firms, and that Problem (CEDEA) has be solved for each of them. In Table 6 we show for each cost function, how often an input has to be changed. For instance, the value 0.09 in the

last row of the Teller column shows that in 9% of all firms, we have to change the number of tellers when the aim is to find a counterfactual instance using the Euclidean norm. When more weight is given to the ℓ_0 norm, few inputs are changed. Indeed, for the Teller column, with the $\ell_0 + \ell_2$, 3% of all firms change it, instead of 9%, and this number decreases to 1% when the $\ell_0 + (\ell_2)$ is used. The same pattern can be observed in all inputs, particularly notable in the Acc. and Ledgers personnel, that goes from changing in more than half of the banks with the Euclidean norm, to changing in only 14% of the firms. The last column of Table 6, Mean $\ell_0(\mathbf{x} - \hat{\mathbf{x}})$, shows how many inputs are changed on average when we use the different cost functions. With the $\ell_0 + (\ell_2)$ only one input has to be decreased, thus with this cost function one detects the crucial input to be modified to be fully efficient, leaving the rest fixed. In general, the results show that, for the inefficient firms, the most common changes leading to full efficiency is to reduce the number of typists and the number of credit officers. The excess of typists is likely related to the institutional setting. Bank branches need the so-called typists for judicial regulations, but they only need the services to a limited degree, see also Table 4. In such cases, it may be difficult to match the full time equivalents employed precisely to the need. The excess of Credit officers is more surprising since, in particular, they are one of the best paid personnel groups.

In Table 7, we look at the size of the changes and not just if a change has to take place or not. The interpretation of the value 0.43 under the first row and the Teller column suggests that when the teller numbers have to be changed, they are reduced by 43% from the initial value, on average. Since several inputs may have to change simultaneously, defining the vector of the relative changes $\boldsymbol{r} = ((x_i - \hat{x}_i)/x_i)_{i=1}^{I}$, the last column shows the mean value of the Euclidean norm of this vector. We see, for example, that in the relatively few cases the teller personnel has to change under $\ell_0 + (\ell_2)$, the changes are relatively large. We see again the difficulties the bank branches apparently have hiring the right amount of typists. We saw in Table 6 that they often have to change and we see now that the changes are non-trivial with about a half excess full time equivalents.

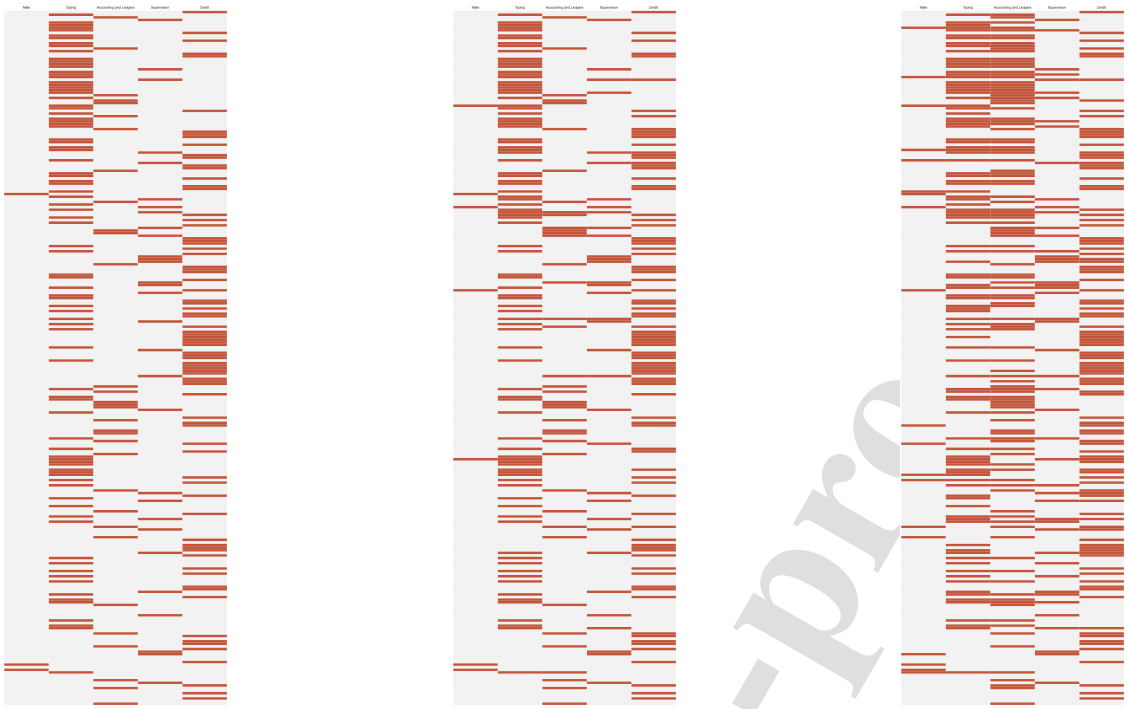
Cost function	Teller	Typing	Acc. and Ledgers	Supervision	Credit	Mean $\ell_0(\mathbf{x} - \hat{\mathbf{x}})$
$\ell_0 + (\ell_2)$	0.01	0.38	0.14	0.13	0.34	1.00
$\ell_0 + \ell_2$	0.03	0.40	0.17	0.14	0.38	1.13
ℓ_2	0.09	0.45	0.51	0.21	0.47	1.72

Table 6: Average results on how often the inputs (personnels) change when desired efficiency is $E^* = 1$.

Cost function	Teller	Typing	Acc. and Ledgers	Supervision	Credit	Mean $\ell_2(\boldsymbol{r})$
$\ell_0 + (\ell_2)$	0.43	0.61	0.41	0.43	0.37	0.4743
$\ell_0 + \ell_2$	0.21	0.58	0.33	0.38	0.35	0.4742
ℓ_2	0.11	0.53	0.14	0.27	0.29	0.4701

Table 7: Average results on how much the inputs (personnels) change when desired efficiency is $E^* = 1$.

In Figure 5, we use a heatmap to illustrate which input factors have to change for the



(a) $C = \ell_0 + (\ell_2)$ (b) $C = \ell_0 + \ell_2$ (c) $C = \ell_2$

Figure 5: The inputs that change when we impose a desired efficiency of $E^* = 1$

individual firms using the three different cost functions in Table 6. Rows with no markings represent firms that were fully efficient to begin with. We see as we would expect that the more weight we put on the Euclidean norm, the more densely populated the illustration becomes, i.e., the more inputs have to change simultaneously.

So far we have asked for counterfactual instances that are fully efficient. If we instead only ask for the counterfactual instances to be at least 80% efficient, only 186 firms need to be studied. As before, let us first visualize the counterfactual explanations for firm 238, which had an original efficiency of $E^0 = 0.72$. In Figure 6, we can see the different changes when imposing $E^* = 0.8$. We again see that the Farrell approach reduces all inputs proportionally, specifically by 9.5% of their values. We see also that under the $\ell_0 + (\ell_2)$ norm, only Credit personnel has to be reduced, by 15%. Under the Euclidean norm, Teller and Acc. and Ledgers personnel are not affected while Typing, Supervision and Credit officers have to be saved, by 4%, 13% and 7%, respectively. Notice that only the change in Supervision is higher in this case than in the Farrell solution, while the decrease in the remain four inputs is significantly smaller for the Euclidean norm. Recall that in Figure 4 the counterfactual explanations for the same firm 238 have been calculated imposing $E^* = 1$. Altering the desired efficiency level from $E^* = 0.8$ to $E^* = 1$ leads to rather dramatic changes in the counterfactual explanations. For the $\ell_0 + (\ell_2)$ cost function, for a desired efficiency of $E^* = 0.8$, we needed to decrease the Credit personnel dramatically

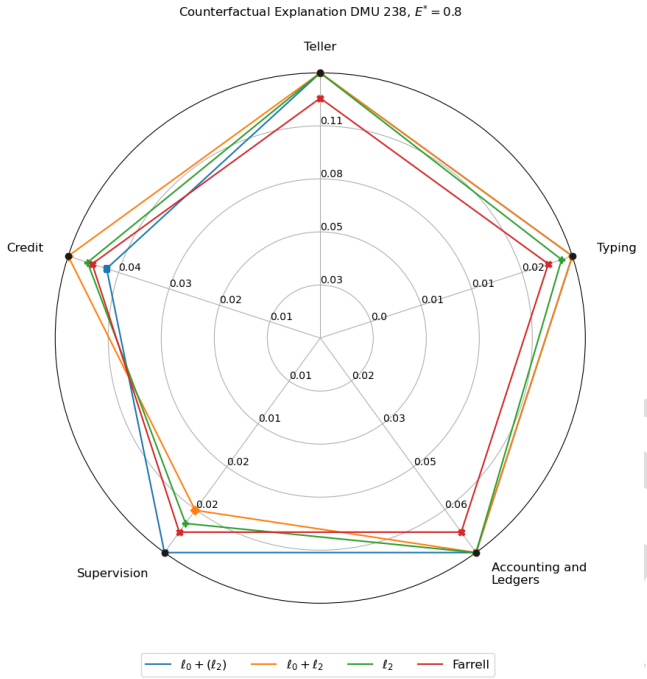


Figure 6: Counterfactual Explanations for DMU 238 with Problem (CEDEA) and desired efficiency $E^* = 0.8$

whereas for a desired efficiency of $E^* = 1$, it is suggested to leave unchanged the Credit personnel and to change the Typing personnel instead. On the other hand, what remains the same is the fact that Teller and Acc. and Ledgers officers are never affected in the counterfactual explanations with the three cost functions.

After the analysis for a single firm, now we present statistics about the counterfactual explanations obtained for all 168 firms that had an original efficiency below 80%. The frequency of changes and the relative sizes of the changes are shown in Tables 8 and 9. We see, as we would expect, that the amount of changes necessary is reduced. On the other hand, the inputs to be changed are not vastly different. The tendency to change in particular credit officers is slightly larger now.

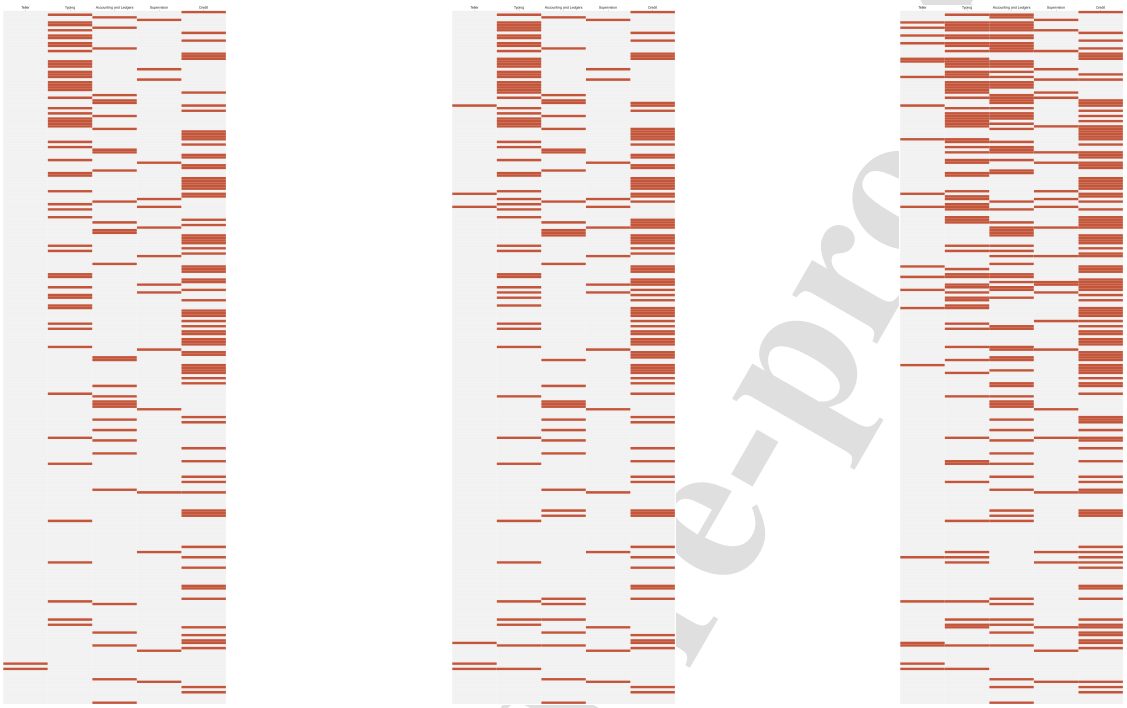
Cost function	Teller	Typing	Acc. and Ledgers	Supervision	Credit	Mean $\ell_0(\mathbf{x} - \hat{\mathbf{x}})$
$\ell_0 + (\ell_2)$	0.01	0.30	0.18	0.08	0.44	1.00
$\ell_0 + \ell_2$	0.03	0.32	0.19	0.09	0.47	1.11
ℓ_2	0.13	0.43	0.51	0.18	0.63	1.88

Table 8: Average results on how often the inputs (personnels) change when desired efficiency is $E^* = 0.8$

In Figure 7, we show the input factors that need to change for the individual firms using the three different cost functions in Table 8 for the case now with $E^* = 0.8$. As expected, we can see now an increasing number of rows with no markings compared to Figure 5, belonging to the firms that had already an efficiency of 0.8.

Cost function	Teller	Typing	Acc. and Ledgers	Supervision	Credit	Mean $\ell_2(\boldsymbol{r})$
$\ell_0 + (\ell_2)$	0.28	0.56	0.28	0.41	0.27	0,3708
$\ell_0 + \ell_2$	0.12	0.52	0.25	0.39	0.26	0.3707
ℓ_2	0.05	0.41	0.11	0.25	0.21	0.3702

Table 9: Average results on how much the inputs (personnels) change when desired efficiency is $E^* = 0.8$



(a) $C = \ell_0 + (\ell_2)$ (b) $C = \ell_0 + \ell_2$ (c) $C = \ell_2$

Figure 7: The inputs that change when we impose a desired efficiency of $E^* = 0.8$

6. Conclusions

In this paper, we have proposed a collection of optimization models to setting targets and finding counterfactual explanations in DEA models, i.e., the least costly changes in the inputs or outputs of a firm that leads to a pre-specified (higher) efficiency level. With our methodology, we are able to include different ways to measure the proximity between a firm and its counterfactual, namely, using the ℓ_0, ℓ_1, and ℓ_2 norms or a combination of them. Calculating counterfactual explanations involves finding "close" alternatives in the complement of a convex set. We have reformulated this bilevel optimization problem as either an MILP or a Mixed Integer Convex Quadratic Problem with linear constraints. In our numerical section, we can see that for our banking application, we are able to solve this model to optimality.

DEA models can capture very complex relationships between multiple inputs and outputs. This allows more substantial evaluations and also offers a framework that can support

many operational, tactical and strategic planning efforts. However, there is also a risk that such a model is seen as a pure black box which in turn can lead to mistrust and some degree of model or algorithm aversion. By looking at counterfactuals, a firm can get a better understanding of the production space and is more likely to trust in the modelling.

Counterfactuals in DEA can also help a firm choose which changes to implement. It is not always enough to simply think of a strategy and which factors can easily be changed, say a direction in input space. It is also important how the technology looks like and therefore how large such changes need to be to get a desired improvement in efficiency. In this way, the analysis of close counterfactuals can help endogenize the choice of both desirable and effective directions to move in. By varying the parameters of the cost function, the firm can even get a menu of counterfactuals, from which it can choose, having thus more flexibility and leading the evaluated firm to gain more trust in the underlying model.

Note also that by calculating the counterfactual explanations for all firms involved, as we did in our banking application, one can determine which combinations of inputs and outputs that most commonly shall be changed to improve efficiency. This is interesting from an overall system point of view. Society at large - or for example a regulator tasked primarily with incentivizing natural monopolies to improve efficiency - may not solely be interested that everyone becomes efficient. It may as well be important how the efficiency is improved, e.g. by reducing the use of imported or domestic resources or by laying off some particular types of labor and not other types.

There are several interesting extensions that can be explored in future research. Here we just mention two. One possibility is to use alternative efficiency measures to constrain the search for counterfactual instances. We have here used Farrell efficiency, which is by far the most common efficiency measure in DEA studies, but one might consider other alternative measures, e.g. additive ones like the excess measure. Another relevant extension could be to make the counterfactuals less individualized. One could for example look for the common features that counterfactual explanations should change across all individual firms and that lead to the minimum total cost.

Acknowledgements

This research has been financed in part by research projects EC H2020 MSCA RISE NeEDS (Grant 822214); COST Action CA19130 - FinAI; FQM-329, P18-FR-2369 and US-1381178 (Junta de Andalucía), PID2019-110886RB-I00 and PID2022-137818OB-I00 (Ministerio de Ciencia, Innovación y Universidades, Spain), and Independent Research Fund Denmark (Grant 9038-00042A) - "Benchmarking-based incentives and regulatory applications". The support is gratefully acknowledged.

References

[1] Agrell, P. and Bogetoft, P. (2017). Theory, techniques and applications of regulatory benchmarking and productivity analysis. In *Oxford Handbook of Productivity Analysis*, pages 523–555. Oxford University Press: Oxford.

[2] Antle, R. and Bogetoft, P. (2019). Mix stickiness under asymmetric cost information. *Management Science*, 65(6):2787–2812.

[3] Aparicio, J., Cordero, J. M., and Pastor, J. T. (2017). The determination of the least distance to the strongly efficient frontier in data envelopment analysis oriented models: modelling and computational aspects. *Omega*, 71:1–10.

[4] Aparicio, J., Mahlberg, B., Pastor, J., and Sahoo, B. (2014). Decomposing technical inefficiency using the principle of least action. *European Journal of Operational Research*, 239:776–785.

[5] Aparicio, J. and Pastor, J. (2013). A well-defined efficiency measure for dealing with closest targets in DEA. *Applied Mathematics and Computation*, 219:9142–9154.

[6] Aparicio, J., Ruiz, J. L., and Sirvent, I. (2007). Closest targets and minimum distance to the pareto-efficient frontier in DEA. *Journal of Productivity Analysis*, 28:209–218.

[7] Bogetoft, P. (2012). *Performance Benchmarking - Measuring and Managing Performance*. Springer, New York.

[8] Bogetoft, P. and Hougaard, J. (1999). Efficiency evaluation based on potential (non-proportional) improvements. *Journal of Productivity Analysis*, 12:233–247.

[9] Bogetoft, P. and Otto, L. (2011). *Benchmarking with DEA, SFA, and R*. Springer, New York.

[10] Carrizosa, E., Ramírez Ayerbe, J., and Romero Morales, D. (2023). Mathematical optimization modelling for group counterfactual explanations. Technical report, IMUS, Sevilla, Spain, https://www.researchgate.net/publication/368958766_Mathematical_Optimization_Modelling_for_Group_Counterfactual_Explanations.

[11] Carrizosa, E., Ramírez Ayerbe, J., and Romero Morales, D. (2024). Generating collective counterfactual explanations in score-based classification via mathematical optimization. *Expert Systems With Applications*, 238:121954.

[12] Charnes, A., Cooper, W. W., Lewin, A. Y., and Seiford, L. M. (1995). *Data Envelopment Analysis: Theory, Methodology and Applications*. Kluwer Academic Publishers, Boston, USA.

[13] Charnes, A., Cooper, W. W., and Rhodes, E. (1978). Measuring the efficiency of decision making units. *European Journal of Operational Research*, 2:429–444.

[14] Charnes, A., Cooper, W. W., and Rhodes, E. (1979). Short Communication: Measuring the Efficiency of Decision Making Units. *European Journal of Operational Research*, 3:339.

[15] Cherchye, L., Rock, B. D., Dierynck, B., Roodhooft, F., and Sabbe, J. (2013). Opening the "black box" of efficiency measurement: Input allocation in multioutput settings. *Operations Research*, 61(5):1148–1165.

[16] Du, M., Liu, N., and Hu, X. (2019). Techniques for interpretable machine learning. *Communications of the ACM*, 63(1):68–77.

[17] Dynan, K. (2000). Habit formation in consumer preferences: Evidence from panel data. *American Economic Review*, 90(3):391–406.

[18] European Commission (2020). *White Paper on Artificial Intelligence : a European approach to excellence and trust.* https://ec.europa.eu/info/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en.

[19] Färe, R. and Grosskopf, S. (2000). Network DEA. *Socio-Economic Planning Sciences*, 34(1):35–49.

[20] Färe, R., Grosskopf, S., and Whittaker, G. (2013). Directional output distance functions: endogenous directions based on exogenous normalization constraints. *Journal of Productivity Analysis*, 40:267–269.

[21] Färe, R., Pasurkac, C., and M.Vardanyan (2017). On endogenizing direction vectors in parametric directional distance function-based models. *European Journal of Operational Research*, 262:361–369.

[22] Fischetti, M. and Jo, J. (2018). Deep neural networks and mixed integer linear optimization. *Constraints*, 23(3):296–309.

[23] Fuhrer, J. C. (2000). Habit formation in consumption and its implications for monetary-policy models. *The American Economic Review*, 90(4):367–390.

[24] Goodman, B. and Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine*, 38(3):50–57.

[25] Guidotti, R. (2022). Counterfactual explanations and how to find them: literature review and benchmarking. Forthcoming in *Data Mining and Knowledge Discovery*.

[26] Gurobi Optimization, L. (2021). Gurobi optimizer reference manual.

[27] Hall, R. (2004). Measuring factor adjustment costs. *The Quarterly Journal of Economics*, 119(3):899–927.

[28] Hamermesh, D. S. and Pfann, G. A. (1999). Adjustment costs in factor demand. *Journal of Economic Literature*, 34(3):1264–1292.

[29] Haney, A. and Pollitt, M. (2009). Efficiency analysis of energy networks: An international survey of regulators. *Energy Policy*, 37(12):5814–5830.

[30] Kao, C. (2009). Efficiency decomposition in network data envelopment analysis: A relational model. *European Journal of Operational Research*, 192:949–962.

[31] Karimi, A.-H., Barthe, G., Schölkopf, B., and Valera, I. (2022). A survey of algorithmic recourse: contrastive explanations and consequential recommendations. *ACM Computing Surveys*, 55(5):1–29.

[32] Lundberg, S. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pages 4765–4774.

[33] Martens, D. and Provost, F. (2014). Explaining data-driven document classifications. *MIS Quarterly*, 38(1):73–99.

[34] Molnar, C., Casalicchio, G., and Bischl, B. (2020). Interpretable machine learning–a brief history, state-of-the-art and challenges. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 417–431. Springer.

[35] Parmentier, A. and Vidal, T. (2021). Optimal counterfactual explanations in tree ensembles. In *International Conference on Machine Learning*, pages 8422–8431. PMLR.

[36] Parmeter, C. and Zelenyuk, V. (2019). Combining the virtues of stochastic frontier and data envelopment analysis. *Operations Research*, 67(6):1628–1658.

[37] Petersen, N. (2018). Directional Distance Functions in DEA with Optimal Endogenous Directions. *Operations Research*, 66(4):1068–1085.

[38] Rigby, D. (2015). Management tools 2015 - an executive's guide. Bain & Company.

[39] Rigby, D. and Bilodeau, B. (2015). Management tools and trends 2015. Bain & Company.

[40] Rostami, S., Neri, F., and Epitropakis, M. (2017). Progressive preference articulation for decision making in multi-objective optimisation problems. *Integrated Computer-Aided Engineering*, 24(4):315–335.

[41] Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., and Zhong, C. (2022). Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistics Surveys*, 16:1–85.

[42] Schaffnit, C., Rosen, D., and Paradi, J. (1997). Best practice analysis of bank branches: An application of DEA in a large Canadian bank. *European Journal of Operational Research*, 98(2):269–289.

[43] Silva Portela, M., Borges, P., and Thanassoulis, E. (2003). Finding closest targets in non-oriented DEA models: The case of convex and non-convex technologies. *Journal of Productivity Analysis*, 19:251–269.

[44] Thach, P. (1988). The design centering problem as a DC programming problem. *Mathematical Programming*, 41(1):229–248.

[45] Wachter, S., Mittelstadt, B., and Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 31:841–887.

[46] Zhu, J. (2016). *Data Envelopment Analysis - A Handbook on Models and Methods*. Springer New York.

[47] Zofio, J. L., Pastor, J. T., and Aparicio, J. (2013). The directional profit efficiency measure: on why profit inefficiency is either technical or allocative. *Journal of Productivity Analysis*, 40:257–266.

Appendix

Here, we extend the analysis in Section 4 by investigating alternative returns to scale and by investigating changes in the outputs rather than the inputs. In both extensions, we will consider a combination of the ℓ_0, ℓ_1 and ℓ_2 norms as in objective function (22).

A. Changing the returns to scale

In Section 4, we have considered the DEA model with constant return to scale (CRS), where the only requirement on the values of $\boldsymbol{\lambda}$ is that they are positive, i.e., $\boldsymbol{\lambda} \in \mathbb{R}_+^{K+1}$, but we could consider other technologies. In that case, to be able to transform our bilevel optimization problem to a single-level one, we should take into account that for each new constraint derived from the conditions on $\boldsymbol{\lambda}$, a new dual variable has to be introduced. We will consider the varying return to scale (VRS) model as it is one of the most preferred one by firms [7], but extensions to other models are analogous.

Consider the input case. With the same transformation as before, we have:

$$\begin{aligned} \min_{\hat{\boldsymbol{x}},F} \quad & C(\boldsymbol{x}^0, \hat{\boldsymbol{x}}) \\ \text{s.t.} \quad & \hat{\boldsymbol{x}} \in \mathbb{R}_+^I \\ & F \leq F^* \end{aligned}$$

$$F \in \underset{\bar{F},\lambda^0,\ldots,\lambda^K}{\arg\min} \{\, \bar{F} : \quad \hat{\boldsymbol{x}} \geq \sum_{k=0}^{K} \beta^k \boldsymbol{x}^k, \hat{F}\boldsymbol{y}^0 \leq \sum_{k=0}^{K} \beta^k \boldsymbol{y}^k, \bar{F} \geq 0, \boldsymbol{\beta} \in \mathbb{R}_+^{K+1}, \sum_{k=0}^{K} \beta^k = F \,\}.$$

Notice that the only difference is that we have a new constraint associated with the technology, namely, $\sum_{k=0}^{K} \beta^k = F$. Let $\kappa \geq 0$ be the new dual variable associated with this constraint. Then, the following changes are made in constraints (19) and (20):

$$\boldsymbol{\gamma}_O^\top \boldsymbol{y}^0 + \kappa = 1 \tag{30}$$

$$\boldsymbol{\gamma}_I^\top \boldsymbol{x}^k - \boldsymbol{\gamma}_O^\top \boldsymbol{y}^k - \kappa \geq 0 \quad k = 0, \ldots, K. \tag{31}$$

The single-level formulation for the counterfactual problem for VRS DEA is as follows:

$$\begin{aligned}
\min_{\hat{\boldsymbol{x}},F,\boldsymbol{\beta},\boldsymbol{\gamma}_I,\boldsymbol{\gamma}_O,\boldsymbol{u},\boldsymbol{v},\boldsymbol{w},\kappa,\boldsymbol{\eta},\boldsymbol{\xi}} \quad & \nu_0 \sum_{i=1}^{I} \xi_i + \nu_1 \sum_{i=1}^{I} \eta_i + \nu_2 \sum_{i=1}^{I} (x_i^0 - \hat{x}_i)^2 \qquad \text{(CEVDEA)} \\
\text{s.t.} \quad & F \leq F^* \\
& \sum_{k=0}^{K} \beta^k = F \\
& \hat{\boldsymbol{x}} \in \mathbb{R}_+^I \\
& \kappa \geq 0 \\
& \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w} \in \{0,1\} \\
& (7) - (10), (15) - (17) \\
& (21), (23) - (31).
\end{aligned}$$

B. Changing the outputs

We have calculated the counterfactual instance of a firm as the mininum cost changes in the inputs in order to have a better efficiency. In the same vein, we could consider instead changes in the outputs, leaving the same inputs. Again, suppose firm 0 is not fully efficient, $E^0 < 1$. Now, we are interested in calculating the minimum changes in the outputs $\boldsymbol{y}^0$ that make it to have a higher efficiency $E^* > E^0$. Let $\hat{\boldsymbol{y}}$ be the new outputs of firm 0 that make it to be at least E^* efficient. We have, then, the following bilevel optimization problem:

$$\begin{aligned}
\min_{\hat{\boldsymbol{y}},E} \quad & C(\boldsymbol{y}^0, \hat{\boldsymbol{y}}) \\
\text{s.t.} \quad & \hat{\boldsymbol{y}} \in \mathbb{R}_+^O \\
& E \geq E^*
\end{aligned}$$

$$E \in \underset{\bar{E},\lambda^0,\ldots,\lambda^K}{\arg\min} \{\, \bar{E} : \quad \bar{E}\boldsymbol{x}^0 \geq \sum_{k=0}^{K} \lambda^k \boldsymbol{x}^k, \hat{\boldsymbol{y}} \leq \sum_{k=0}^{K} \lambda^k \boldsymbol{y}^k, \bar{E} \geq 0, \boldsymbol{\lambda} \in \mathbb{R}_+^{K+1} \,\}.$$

Following similar steps as in previous section, the single-level formulation for the counterfactual problem in DEA in the output case is as follows:

$$\min_{\hat{\boldsymbol{y}},E,\boldsymbol{\lambda},\gamma_I,\gamma_O,\boldsymbol{u},\boldsymbol{v},\boldsymbol{w},\boldsymbol{\eta},\boldsymbol{\xi}} \quad \nu_0 \sum_{o=1}^{O} \xi_o + \nu_1 \sum_{o=1}^{O} \eta_o + \nu_2 \sum_{o=1}^{O} (y_o^0 - \hat{y}_o)^2 \qquad \text{(CEODEA)}$$

$$\begin{aligned}
\text{s.t.} \quad & \hat{\boldsymbol{y}} \in \mathbb{R}_+^O \\
& E \geq E^* \\
& E\boldsymbol{x}^0 \geq \sum_{k=0}^{K} \lambda^k \boldsymbol{x}^k \\
& \hat{\boldsymbol{y}} \leq \sum_{k=0}^{K} \lambda^k \boldsymbol{y}^k \\
& \boldsymbol{\gamma}_I^\top \boldsymbol{x}^0 = 1 \\
& \boldsymbol{\gamma}_O^\top \boldsymbol{y}^k - \boldsymbol{\gamma}_I^\top \boldsymbol{x}^k \leq 0 && k = 0,\ldots,K \\
& \gamma_I^i \leq M_I u_i && i = 1,\ldots,I \\
& E x_i^0 - \sum_{k=0}^{K} \lambda^k x_i^k \leq M_I(1-u_i) && i = 1,\ldots,I \\
& \gamma_O^o \leq M_O v_o && o = 1,\ldots,O \\
& -\hat{y}_o + \sum_{k=0}^{K} \lambda^k y_o^k \leq M_O(1-v_o) && o = 1,\ldots,O \\
& \lambda^k \leq M_f w_k && k = 0,\ldots,K \\
& \boldsymbol{\gamma}_O^\top \boldsymbol{y}^k - \boldsymbol{\gamma}_I^\top \boldsymbol{x}^k \leq M_f(1-w_k) && k = 0,\ldots,K \\
& -M_{\text{zero}}\xi_o \leq y_o^0 - \hat{y}_o && o = 1,\ldots,O \\
& y_o^0 - \hat{y}_o \leq M_{\text{zero}}\xi_o && o = 1,\ldots,O \\
& \eta_o \geq y_o^0 - \hat{y}_o && o = 1,\ldots,O \\
& -\eta_o \leq y_o^0 - \hat{y}_o && o = 1,\ldots,O \\
& E, \boldsymbol{\lambda}, \boldsymbol{\gamma}_I, \boldsymbol{\gamma}_O, \boldsymbol{\eta} \geq 0 \\
& \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}, \boldsymbol{\xi} \in \{0,1\}.
\end{aligned}$$

As in the input model, depending on the cost function, we either obtain an MILP model or a Mixed Integer Convex Quadratic model with linear constraints. This model could be formulated analogously for the VRS case.

Counterfactual Analysis and Target Setting in Benchmarking

- Peter Bogetoft

  Department of Economics, Copenhagen Business School, Frederiksberg, Denmark

  pb.eco@cbs.dk

- Jasone Ramírez-Ayerbe

  Instituto de Matemáticas de la Universidad de Sevilla, Seville, Spain

  mrayerbe@us.es

- Dolores Romero Morales

  Department of Economics, Copenhagen Business School, Frederiksberg, Denmark

  drm.eco@cbs.dk

Corresponding Author:
Jasone Ramírez-Ayerbe
Instituto de Matemáticas de la Universidad de Sevilla, Seville, Spain
Tel: (+34) 699421303
Email: mrayerbe@us.es

Acknowledgements

This research has been financed in part by research projects EC H2020 MSCA RISE NeEDS (Grant 822214); COST Action CA19130 - FinAI; FQM-329, P18-FR-2369 and US-1381178 (Junta de Andalucía), PID2019-110886RB-I00 and PID2022-137818OB-I00 (Ministerio de Ciencia, Innovación y Universidades, Spain), and Independent Research Fund Denmark (Grant 9038-00042A) - "Benchmarking-based incentives and regulatory applications". The support is gratefully acknowledged.

Highlights

- Counterfactual analysis to set targets in Data Envelopment Analysis
- Counterfactuals improve performance in the easiest possible way
- A novel bilevel optimization model to find counterfactuals
- Ability to control desired efficiency and number of input/outputs moved