

Journal of Mathematical Extension
Vol. 20, No. 4 (2026) (4) 1-27
ISSN: 1735-8299
URL: <http://doi.org/10.30495/JME.2026.3437>
Original Research Paper

Cost Congestion in Data Envelopment Analysis

S. Asadi Rahmati *

Technical and Vocational University (TVU)

Abstract.

Identifying the sources of inefficiency is essential for accurately evaluating the performance of Decision Making Units (DMUs) in Data Envelopment Analysis (DEA). While congestion has been extensively studied as a major source of inefficiency, existing approaches primarily focus on input-output relationships and overlook inefficiencies associated with the cost structure. To address this gap, this study introduces the novel concept of cost congestion, defined as a situation in which an increase in one or more cost components reduces output, or conversely, a reduction in cost components improves output. Building upon this concept, the proposed framework provides a comprehensive analysis of cost-related inefficiencies by distinguishing the effects of cost congestion, technical inefficiencies arising from input utilization, and inefficiencies associated with input prices. Furthermore, the study develops a decomposition approach to quantify lost opportunities attributable to each source of inefficiency, enabling a more detailed assessment of performance. The proposed methodology extends the conventional DEA framework by integrating cost-based congestion analysis and offers managers and decision-makers a more informative tool for identifying inefficiency sources and prioritizing corrective actions to improve operational performance.

AMS Subject Classification: 90B50, 62P20

Keywords and Phrases: Cost congestion, cost efficiency, production possibility set, decomposition of efficiency

Received: July 2025; Accepted: June 2026

*Corresponding Author

1 Introduction

Data Envelopment Analysis (DEA), initially introduced by Charnes, Cooper, and Rhodes (1978) [3], is a method for evaluating the performance of similar Decision Making Units (DMUs) that use one or more inputs to generate one or more outputs. It is generally assumed that increasing inputs will lead to an increase in outputs. As inputs increase, a corresponding rise in outputs is typically expected. However, in some cases, an increase in inputs is associated with a decrease in outputs, which can lead to reduced revenue and profits. This counterintuitive phenomenon is known as congestion. For example, an increase in the number of workers in a mine may lead to overcrowding, which can reduce efficiency rather than enhance it. Generally, a DMU is in congestion if the reduction of one or more inputs accompanied by the increase in one or more outputs without worsening the other outputs, or conversely, the increase in one or more inputs is accompanied by the reduce in one or more outputs without worsening the other outputs. Various methods have been introduced to examine the congestion.

Congestion was first defined by [15] as a consequence of the law of variable proportions, that plays a vital role in measuring production performance and managing energy consumption in farming [18]. Subsequently, [11] conceptualized congestion as a form of technical inefficiency that arises independently of price information. In [8], a slack-based approach was proposed, wherein congestion is defined as the deviation between observed and expected input-output quantities. [1] investigated inefficiencies and congestion in the Chinese industrial sector before and after the 1978 economic reforms, demonstrating how congestion models can be used to identify trades that enhance output and labor efficiency. Moreover, their method can be applied in situations where increasing certain inputs is beneficial, much like the increase in outputs [7] established necessary and sufficient conditions for the presence of congestion and offered new interpretations of slack-based Data Envelopment Analysis (DEA), incorporating these insights into their extended framework. In response, [13] provided comments on the work of Cooper et al. [7], arguing that their approach detects congestion when the strong efficiency projection contains non-zero slacks, although the reverse implication does not necessarily hold. Navidi et al. [21] Proposed a method

that offers a model-free framework for measuring congestion, efficiency, and inefficiency, capable of handling diverse production sets and data types while distinguishing between strong and weak congestion levels in DMUs. Unlike previous models that often struggle with computational complexity and limited spatial coverage, the method proposed by Salehi et al. [23] leverages fundamental definitions to identify congestion and its severity. This approach not only ensures high precision but also provides a seamless identification across all units. Karimi et al. [17] extended the application of MIP models to integer-constrained environments, addressing the challenge of identifying and eliminating congestion when inputs and outputs are limited to integer values, while ensuring the integration of an integer reference point. Haghighi et al. [16] introduced revenue congestion as a state where input reductions drive revenue growth via output rebalancing. This paper further develops this concept by proposing a method to systematically identify and quantify its sources. Several studies have reviewed and compared existing approaches for estimating congestion in DEA and quantifying the associated output losses [26]. Among the available methods, the approach proposed by [20] is particularly attractive because of its conceptual simplicity and low computational burden. Moreover, it shares many characteristics with the method developed by [5]. Owing to these advantages, the present study adopts the framework of [20] for congestion identification. The existing DEA literature consistently recognizes congestion as an important source of technical inefficiency, and eliminating congestion is generally expected to improve production performance. Nevertheless, despite the extensive body of research on congestion, its role as a source of cost inefficiency has received little attention. Traditional congestion analysis focuses on the relationship between physical inputs and outputs, where excessive input usage may reduce output levels. From a cost perspective, however, a similar phenomenon may occur when excessive expenditure on one or more cost components results in lower output levels, or when reducing certain cost components leads to improved outputs. This observation motivates the introduction of the concept of cost congestion. Although cost efficiency has been extensively investigated in the DEA literature, the influence of cost congestion on overall cost inefficiency has not yet been explicitly incorporated into existing decomposition frame-

works. The decomposition of cost efficiency has been widely studied in DEA. [24] demonstrated that, under the assumptions of a single input and identical input prices, the conventional cost efficiency models of [10], [4], [2], and [6] produce results that are equivalent to technical efficiency. Furthermore, [24] showed that two DMUs with identical input-output levels may exhibit identical cost efficiency scores even when their input and output prices differ substantially, thereby revealing an inconsistency in conventional cost efficiency models. To overcome this limitation, a new cost-based production possibility set was proposed. Subsequently, [9] developed an advanced decomposition of cost inefficiency for noncompetitive environments in which input prices vary across DMUs. Their framework introduced separate price and cost production possibility sets and decomposed cost inefficiency into technical, price, and congestion-related components. Similarly, [25] decomposed cost efficiency into technical, price, and allocative efficiency. Despite these significant contributions, the literature has not yet provided a comprehensive framework that explicitly incorporates cost congestion into cost efficiency analysis. In particular, the interaction between congestion and cost inefficiency has not been adequately investigated, especially in production environments characterized by heterogeneous input prices. Ignoring this source of inefficiency may lead to an incomplete understanding of performance losses and consequently to suboptimal managerial decisions. To address this research gap, the present study extends the conventional DEA framework by introducing the concept of cost congestion and integrating it into the decomposition of cost inefficiency. The proposed methodology distinguishes inefficiencies arising from cost congestion, technical inefficiency associated with input utilization, and inefficiency due to input prices. Furthermore, it quantifies the opportunity loss attributable to each source of inefficiency, thereby providing a more comprehensive assessment of cost performance. The effectiveness of the proposed approach is illustrated through both a numerical example and an empirical application, demonstrating its practical value for performance evaluation and managerial decision-making.

2 Congestion

Congestion in economics refers to a situation where an increase in one or more inputs leads to a decrease in one or more outputs without improving other inputs or outputs. Fare and Svensson [15] defined congestion in relation to the law of variable proportions, while Fare and Grosskopf [14] proposed methods to identify the specific inputs responsible for congestion. Brockett et al. [1] and Cooper et al. [7] developed DEA-based approaches to measure input congestion. Among the existing approaches, the model of Fare and Svensson [15] and Cooper et al. [7] are considered foundational. However, the Cooper et al. method can involve considerable computational complexity. To address this issue, this study adopts the method proposed by Noura et al. [20], which provides a computationally simplified framework for measuring congestion while producing results equivalent to those of the Cooper et al. model.

Consider n DMU_j ($j = 1, \dots, n$), which uses m inputs x_{ij} ($i = 1, \dots, m$) to generate s outputs y_{rj} ($r = 1, \dots, s$). Suppose the production possibility set is defined as follows [20]:

$$P = \left\{ (X, Y) \mid X \geq \sum_{j=1}^n \lambda_j X_j, Y \leq \sum_{j=1}^n \lambda_j Y_j, \sum_{j=1}^n \lambda_j = 1, \lambda_j \geq 0, \right. \\ \left. j = 1, \dots, n \right\}.$$

Method is used to examine the congestion. The following model is first

solved in order to examine the congestion of DMU_o :

$$\begin{aligned}
& \max \varphi_o + \varepsilon \left(\sum_{i=1}^m s_i^- + \sum_{r=1}^s s_r^+ \right) \\
\text{s.t. } & \sum_{j=1}^n \lambda_j x_{ij} + s_i^- = x_{io} \quad , i = 1, \dots, m \\
& \sum_{j=1}^n \lambda_j y_{rj} - s_r^+ = \varphi_o y_{ro} \quad , r = 1, \dots, s \\
& \sum_{j=1}^n \lambda_j = 1 \\
& \lambda_j \geq 0 \quad , j = 1, \dots, n \\
& s_i^- \geq 0, s_r^+ \geq 0 \quad , \text{for all } i, r.
\end{aligned} \tag{1}$$

The set E is considered to be a set of efficient units as follows:

$$E = \{j \mid \varphi_j = 1\}.$$

Now, from the set of efficient units E , select those units that have the highest value in at least one input. The corresponding input for each selected unit is denoted by $x_i^{\max}$.

Congestion exists at DMU_o if and only if at least one of the following conditions is satisfied in the optimal solution to model (1).

- (I) $\varphi_o^* > 1$, and at least in one component $x_{io} > x_i^{\max}$.
- (II) At least in one component $s_r^{+*} > 0$ and at least in one component $x_{io} > x_i^{\max}$.

In this case, the congestion value in i -th input is $S_i^c = x_{io} - x_i^{\max}$.

This section expands on cost inefficiency by incorporating variations in input prices across units. It introduces new sources of inefficiency, including congestion effects related to both inputs and prices, within an extended production possibility framework.

3 Cost Inefficiency and the Introduction of Its New Resources

Fare et al. [10], proposed the following model for calculating the cost efficiency:

$$\begin{aligned} cx^* &= \min_{x, \lambda} cx \\ x &\geq X\lambda \\ y_o &\leq Y\lambda \\ \lambda &\geq 0. \end{aligned}$$

This model can be used with the assumption of the same price data for all units. In case where we have n inputs $X = (x_1, \dots, x_n)$. Suppose that DMUs A and B have the same amount of inputs and outputs, i.e., $x_A = x_B$ and $y_A = y_B$. Assume further that the unit cost of DMU A is twice that of DMU B for each input, i.e., $C_A = 2C_B$. Under these assumptions, Both DMUs A and B have the same cost (overall) and allocative efficiencies. These shortcomings are caused by the structure of the supposed production possibility set P as defined by:

$$P = \{(x, y) | x \geq \bar{X}, y \leq \bar{Y}, X \geq 0\}$$

P is defined only by using technical factors $X = (x_1, \dots, x_n)$ and $Y = (y_1, \dots, y_n)$, but has no concern with the unit input cost $C = (c_1, \dots, c_n)$. Let us define another cost-based production possibility set P_c as:

$$\begin{aligned} \bar{P}_c &= \left\{ (\bar{X}, Y) \mid \bar{X} \geq \sum_{j=1}^n \lambda_j \bar{X}_j, Y \leq \sum_{j=1}^n \lambda_j Y_j, \sum_{j=1}^n \lambda_j = 1, \lambda_j \geq 0, \right. \\ &\quad \left. j = 1, \dots, n \right\}, \end{aligned}$$

where $(\bar{x}_{ij} = (c_{ij}x_{ij}))$. In this production possibility set, the performance of the units was evaluated. To evaluate cost-related performance, it is essential to first establish a conceptual framework for the production environment. Conceptually, the cost production possibility set, denoted

by $\bar{P}_c$, represents the collection of all feasible output levels and their corresponding minimum required expenditures, given the prevailing input prices and available technology. Rather than focusing solely on physical input quantities, this set provides an economic benchmark by identifying the lowest cost at which a specific output bundle can be produced. It serves as the foundation for measuring cost inefficiency, as it allows us to distinguish between units that are operating on the cost frontier and those that are incurring avoidable expenses. [25] provides a clear breakdown of cost inefficiency and analyzes the cost losses caused by its different components. In addition to the above factors, this section introduces the factors related to the congestion caused by inputs and the congestion caused by prices. Based on [24], the production possibility set $\bar{P}_c$ corresponding to the data in Table 2 is depicted in Figure 1.

Figure 1: Cost Production Possibility Set ($\bar{P}_c$)

As illustrated, units $\bar{B}$, $\bar{E}$, and $\bar{F}$ lie on the frontier of $\bar{P}_c$. Among them, only unit $\bar{B}$ is cost-efficient, as it achieves the lowest cost value measured along the horizontal axis, which represents the product of input quantities and their corresponding prices. In contrast, unit $\bar{E}$, despite being on the frontier, is allocatively inefficient. Cost is a function of input quantities and their corresponding prices; therefore, any inefficiency in cost is inherently linked to these components. [25] initially identified input-related inefficiency using a DEA model within the conventional production possibility set P . Subsequently, by analyzing the points formed through the multiplication of input prices and the projected input quantities in the space $\bar{P}_c$, they identified additional sources of inefficiency, specifically price inefficiency and allocative inefficiency.

Figure 2: Production Possibility Sets Made by Inputs and Outputs (P)

Figure 2 shows the conventional production possibility set P for the data in Table 2. Units A , B , C , D and E are the technical efficient, and other units are technically inefficient. Unit G is a unit that has a congestion in the input value, since, by reducing the input value to one unit, its output value is not reduced. This congestion will also cause cost

inefficiency. Thus, congestion is also considered to be one of the most important factors in cost efficiency, and this is considerable. Based on the set $\bar{P}_c$ in Figure 1, units $\bar{C}$, $\bar{D}$ and $\bar{G}$ are in the congestion region and, because they are in the cost production possibility set, they are said to be in a cost congestion position. The current position of unit $\bar{G}$ in Figure 2 is related to the time when the congestion has been resolved in the input value and technical inefficiency related to unit G in space P , and therefore, the remaining congestion in the cost space $\bar{P}_c$ is related to the input price.

Given the observed cost of the initial point G (the value of G multiplied by the cost G) in space $\bar{P}_c$ (Figure 1) it can be observed that one gets to the size of an input unit to the point G^c by removing the input congestion, and then, one gets to the point $\bar{G}$ by removing the technical inefficiency, which is in a cost space in the congestion position. Point $\bar{G}^c$ is obtained, after removing cost congestion. Ultimately, the projection into the $\bar{P}_c$'s frontier is point $\bar{G}^*$. Therefore, it can be seen that any inefficiencies of the type of cost congestion can be caused by input congestion and technical input inefficiency.

Moreover a unit in conventional space P might have a congestion, but its corresponding point might not have the congestion in the space $\bar{P}_c$. In addition, a unit such as a $\bar{C}$ may have a cost congestion (Figure 1), and its corresponding point (C point) in P might not have any input congestion or input technical inefficiency (Figure 2), and the result cost congestion might be only due to the input price. Therefore, cost congestion can be due to congestion in input, input technical inefficiency and input price, and cost technical inefficiency can be due to input congestion, technical input inefficiency, and input prices. For the analysis of these conditions, the next section will analyze the cost efficiency into different segments.

4 Decompose Cost Efficiency into Distinct Components

This section systematically decomposes overall cost efficiency into its fundamental components. By analyzing observed input costs, congestion effects, technical inefficiency, price-related inefficiencies, and alloca-

tive inefficiency, it offers a comprehensive framework to understand cost losses in DEA.

This subsection defines the observed input cost as the product of input quantities and their corresponding prices for the given DMU, serving as the baseline for further analysis.

4.1 Observed input cost

If $x_o = (x_{1o}, x_{2o}, \dots, x_{mo})$ be the inputs, $y_o = (y_{1o}, y_{2o}, \dots, y_{so})$ the outputs, and $c_o = (c_{1o}, c_{2o}, \dots, c_{mo})$ the prices related to each input unit for the observed DMU (DMU_o), then the actual observed cost will be obtained by the observed input multiplication at the observed price, as follows:

$$C_o = \sum_{i=1}^m c_{io} x_{io}.$$

Here, congestion is identified as the inefficiency arising from excessive input usage, which can be reduced to improve cost efficiency by eliminating redundant inputs.

4.2 Loss caused by congestion

With the assumption of a conventional production possibility set P , the congestion in (DMU_o) is examined by means of Model (1). If there is a congestion at each input, the amount of congestion of that input will be decreased from that input value.

$$x_{io}^c = x_{io} - s_i^c.$$

The cost in a unit without a congestion x_o^c will obtain as follows:

$$C_o^* = \sum_{i=1}^m c_{io} x_{io}^c.$$

As $x_{io}^c < x_{io}$, the following consequences arise:

$$C_o^* = \sum_{i=1}^m c_{io} x_{io}^c < \sum_{i=1}^m c_{io} x_{io} = C_o.$$

Congestion efficiency is defined as follows:

$$\frac{C_o^*}{C_o} = \frac{\sum_{i=1}^m c_{io} x_{io}^c}{\sum_{i=1}^m c_{io} x_{io}} = \frac{\sum_{i=1}^m c_{io} (x_{io} - s_i^c)}{\sum_{i=1}^m c_{io} x_{io}}.$$

In this case, the difference between these two costs is defined as the loss caused by the input congestion, as follows:

$$L_o^* = C_o - C_o^*.$$

This part evaluates inefficiencies related to the production technology, measuring how far the DMU is from operating on the efficient frontier without congestion.

4.3 Loss caused by technical inefficiency

A suitable model (CCR, BCC, SBM, or ...) is used to verify the technical efficiency of a without congestion unit (x_o^c) in space P :

$$\begin{aligned} & \min \theta_o - \varepsilon \left(\sum_{i=1}^m s_i^- + \sum_{r=1}^s s_r^+ \right) \\ \text{s.t. } & \sum_{j=1}^n \lambda_j x_{ij} + s_i^- = \theta_o x_{io}^c, \quad i = 1, \dots, m \\ & \sum_{j=1}^n \lambda_j y_{rj} - s_r^+ = y_{ro}, \quad r = 1, \dots, s \\ & \sum_{j=1}^n \lambda_j = 1 \\ & \lambda_j \geq 0, \quad j = 1, \dots, n \\ & s_i^- \geq 0, \quad s_r^+ \geq 0, \quad \text{for all } i, r. \end{aligned} \tag{2}$$

If $(\theta^*, \lambda^*, s^{-*}, s^{+*})$ be the optimal solution of Model (2), then the efficient input (x_0^*) will be as follows:

$$x_o^* = \theta^* x_o^c - s^{-*}.$$

The technical efficient input cost for the DMU_o are presented below:

$$C_o^{**} = \sum_{i=1}^m c_{io} x_{io}^*.$$

The technical efficiency is given by:

$$\frac{C_o^{**}}{C_o^*} = \frac{\sum_{i=1}^m c_{io} x_{io}^*}{\sum_{i=1}^m c_{io} x_{io}^c} = \frac{\sum_{i=1}^m c_{io} (\theta^* x_{io}^c - s_i^{-*})}{\sum_{i=1}^m c_{io} x_{io}^c},$$

where

$$C_o^{**} = \sum_{i=1}^m c_{io} x_{io}^* = \sum_{i=1}^m c_{io} (\theta^* x_{io}^c - s_i^{-*}) < \sum_{i=1}^m c_{io} \theta^* x_{io}^c < \sum_{i=1}^m c_{io} x_{io}^c = C_o^*.$$

Therefore, the cost loss caused by technical inefficiency is defined as:

$$L_o^{**} = C_o^* - C_o^{**}.$$

This subsection examines congestion effects in the cost space, where inefficient input pricing leads to higher costs despite optimal input usage.

4.4 Loss caused by price congestion

The cost production possibility set $\bar{P}_c$ is considered, and all the points depicted in the previous step are moved into space $\bar{P}_c$, by multiplying the cost related to that unit ($\bar{x}_{io}^* = c_{io} x_{io}^*$). Now the existence of the price congestion of $\bar{x}_{io}^*$ in $\bar{P}_c$ as detailed below:

$$\begin{aligned} & \min \bar{\varphi}_o - \varepsilon \left(\sum_{i=1}^m \bar{s}_i^- + \sum_{r=1}^s \bar{s}_r^+ \right) \\ \text{s.t. } & \sum_{j=1}^n \lambda_j \bar{x}_{ij} + \bar{s}_i^- = \bar{x}_{io}^*, \quad i = 1, \dots, m \\ & \sum_{j=1}^n \lambda_j y_{rj} - \bar{s}_r^+ = \bar{\varphi}_o y_{ro}, \quad r = 1, \dots, s \\ & \sum_{j=1}^n \lambda_j = 1, \quad \lambda_j \geq 0, \quad j = 1, \dots, n \\ & s_i^- \geq 0, \quad s_r^+ \geq 0, \quad \text{for all } i, r. \end{aligned} \tag{3}$$

The set $\overline{E}$ is considered in the form of a set of efficient units can be stated as:

$$\overline{E} = \{j | \overline{\varphi}_j^* = 1\}.$$

Now, among the efficient units of the set $\overline{E}$, the units that have the highest value in at least one input will be selected, and the input related to that unit will be specified as $\overline{x}_i^{\max}$. At DMU_o , there is a congestion if and only if at least one of the following conditions is met in the optimal solution of the Model (1).

1. $\overline{\varphi}_o^* > 1$ and at least in one component $\overline{x}_{io} > \overline{x}_i^{\max}$.
2. At least in one component $\overline{s}_r^{+*} > 0$ and at least in one component $\overline{x}_{io} > \overline{x}_i^{\max}$.

In this case, the value of i-th input price congestion is as $\overline{S}_i^c = \overline{x}_{io} - \overline{x}_i^{\max}$. In the case of a congestion, the input without the cost congestion can be written as:

$$\overline{x}_{io}^c = \overline{x}_{io} - \overline{s}_i^c.$$

The input cost of point without a congestion $\overline{x}_o^c$ is obtained as follows:

$$C_o^{***} = \sum_{i=1}^m \overline{x}_{io}^c,$$

where

$$C_o^{***} = \sum_{i=1}^m \overline{x}_{io}^c = \sum_{i=1}^m (\overline{x}_{io} - \overline{s}_i^c) < \sum_{i=1}^m \overline{x}_{io} = C_o^{**}.$$

The price congestionall efficiency is formulated as:

$$\begin{aligned} \frac{C_o^{***}}{C_o^{**}} &= \frac{\sum_{i=1}^m c_{io} \overline{x}_{io}^c}{\sum_{i=1}^m c_{io}} \\ &= \frac{\sum_{i=1}^m \overline{x}_{io}^c}{\sum_{i=1}^m c_{io} (\theta^* \overline{x}_{io} - \overline{s}_i^{-*})} \\ &= \frac{\sum_{i=1}^m \overline{x}_{io}^c}{\sum_{i=1}^m (\theta^* \overline{x}_{io}^c - c_{io} \overline{s}_i^{-*})}. \end{aligned}$$

The loss of cost caused by the existence of congestion in the cost space $\bar{P}_c$ is defined as follows:

$$L_o^{***} = C_o^{**} - C_o^{***}.$$

Here, the focus is on technical inefficiencies specifically related to input prices, highlighting further opportunities to reduce costs.

4.5 Loss caused by input price inefficiency

Now at the point $\bar{x}_o^c$, whose congestion is removed, the price technical inefficiency is examined. In this way, the BCC model is used in the cost space $\bar{P}_c$:

$$\begin{aligned} & \min \rho_o - \varepsilon \left(\sum_{i=1}^m t_i^- + \sum_{r=1}^s t_r^+ \right) \\ \text{s.t. } & \sum_{j=1}^n \mu_j \bar{x}_{ij} + t_i^- = \rho_o \bar{x}_{io}^c, \quad i = 1, \dots, m \\ & \sum_{j=1}^n \mu_j y_{rj} - t_r^+ = y_{ro}, \quad r = 1, \dots, s \\ & \sum_{j=1}^n \mu_j = 1, \\ & \mu_j \geq 0, \quad j = 1, \dots, n \\ & t_i^- \geq 0, \quad t_r^+ \geq 0, \quad \text{for all } i, r. \end{aligned} \quad (4)$$

If $(\rho^*, \mu^*, t^{*-}, t^{+*})$ be the optimal solution of (4), the efficient input x_0^* will be as follows:

$$\bar{x}_o^* = \mu^* \bar{x}_o^c - t^{*-}.$$

And the technical efficiency input price is as follows:

$$C_o^{****} = \sum_{i=1}^m \bar{x}_{io}^* = \sum_{i=1}^m \mu^* \bar{x}_{io}^c - t^{*-} < \sum_{i=1}^m \mu^* \bar{x}_{io}^c < \sum_{i=1}^m \bar{x}_{io}^c = C_o^{***}.$$

The price technical efficiency is defined as follows:

$$\frac{C_o^{****}}{C_o^{***}} = \frac{\sum_{i=1}^m \mu^* \bar{x}_{io}^c - t^{*-}}{\sum_{i=1}^m \bar{x}_{io}^c}.$$

And the loss caused by technical inefficiency in the cost space $\bar{P}_c$ is defined as follows:

$$L_o^{****} = C_o^{***} - C_o^{*****}.$$

This part considers inefficiencies due to non-optimal allocation of input prices, reflecting potential cost savings by adjusting input price combinations.

4.6 Loss caused by allocative inefficiency

Now, the following cost model is solved in the cost space $\bar{P}_c$ to find the point with the lowest cost:

$$\begin{aligned} C_o^{*****} &= \min e\bar{x} \\ \text{s.t. } \sum_{j=1}^n \mu_j \bar{x}_{ij} + t_i^- &= \bar{x}_i, \quad i = 1, \dots, m \\ \sum_{j=1}^n \mu_j y_{rj} - t_r^+ &= y_{ro}, \quad r = 1, \dots, s \\ \sum_{j=1}^n \mu_j &= 1 \\ \mu_j &\geq 0, \quad j = 1, \dots, n \\ t_i^- \geq 0, \quad t_r^+ &\geq 0, \quad \text{for all } i, r. \end{aligned} \tag{5}$$

If (x_o^*, μ_0^*) be the optimal solution of the Model (5), the cost pair $(\bar{x}_o, y_o)$ in $\bar{P}_c$ will have the lowest cost among all units. Therefore, the allocative efficiency caused by inappropriate prices, compared to the best price, may be defined as follows:

$$(\text{Allocative Efficiency}) \text{ AE} = \frac{C_o^{*****}}{C_o^{****}}.$$

The loss caused by the nonoptimal combination of the price, the allocative inefficiency, is defined as follows:

$$L_o^{*****} = C_o^{****} - C_o^{*****}.$$

Finally, the section presents a comprehensive decomposition formula, showing how total observed cost can be broken down into the sum of lowest achievable cost and various inefficiency losses.

4.7 The composition of inefficiency and costs

The cost (overall) efficiency, the ratio of the lowest possible cost combination (C_o^{*****}) to the observed cost (C_o), can be broken down into the above mentioned factors:

$$\begin{aligned} \text{Overall Eff.} = & \text{Allocative Eff.} \times \text{Technical Eff.} \times \text{Price Cogestionall Eff.} \\ & \times \text{Price Technical Eff.} \times \text{Congestionall Eff.}, \end{aligned}$$

which implies:

$$\frac{C_o^{*****}}{C_o} = \frac{C_o^{*****}}{C_o^{****}} \times \frac{C_o^{****}}{C_o^{***}} \times \frac{C_o^{***}}{C_o^{**}} \times \frac{C_o^{**}}{C_o^*} \times \frac{C_o^*}{C_o}.$$

Moreover the actual observed cost (C_o) can be stated as the sum of the lowest cost and the loss caused by congestion, technical inefficiency, price congestion inefficiency and price technical inefficiency and allocation inefficiency:

$$C_o = L_o^{*****} + L_o^{****} + L_o^{***} + L_o^{**} + L_o^* + C_o^{*****}.$$

This section applies the proposed methodology to numerical data, illustrating how cost congestion and inefficiencies can be identified and quantified. The examples demonstrate the practical relevance of the cost efficiency decomposition framework.

Table 1: Summary of Cost Components

Observed input cost	c_o
Unit with out congestion cost	c_o^*
The technicale efficient input cost	c_o^{**}
The input cost of $\bar{x}_o^c$ (unit without conestion)	c_o^{***}
Technical efficiency input price	c_o^{****}
The lowest cost in the cost space $\bar{P}_c$	c_o^{*****}

5 Numerical Examples

A simple example with eight Decision Making Units (DMUs) is used to explain the detection and elimination of congestion in both input

and cost spaces. The decomposition of cost efficiency into congestion, technical, price, and allocative inefficiencies is clearly demonstrated.

5.1 An illustrative example

This section explains the proposed method and its related decompositions about the data in Table 2, in more details.

Consider a set of 8 DMUs with an input and an output. Table 2 shows the input and output values in the second and third columns, and the price of each input unit in the fourth column. In addition, it shows the observed cost for units in column 6.

Table 2: Data of 8 DMUs

DMU	Input	Output	Price of Input (c_o)	Observed input cost ($C_j = c_j x_j$)	φ_o
A	1.5	1	2	3	1
B	2	2	1	2	1
C	3	3	5	15	1
D	4	4	3	12	1
E	6	5	1.5	9	1
F	7	5	1.75	12.25	1
G	8	4	3	24	1.25
H	6	3	2	12	1.67

By means of Model (1), one can clearly see that unit G is in congestion. Table 2 shows the amount of efficiencies by means of Model (1). Units A, B, C, D, E, F are efficient, that is, $E = \{A, B, C, D, E, F\}$. Among these units, unit F has the highest input value $x_1^{\max} = 7$. Unit G is inefficient as $8 = x_1^G > x_1^{\max}$ and $\varphi^G > 1$. Therefore, unit G is in congestion, and its congestion value is $S^c = 8 - 7 = 1$.

Once the congestion is removed, point G^c is obtained (Figure 2). The technical efficiency value of the units lacking congestion, has been obtained by means of the BCC model (Column 6, Table 3). The technical projection of the inputs has been presented in the seventh column of Table 3. Technical projection of unit G is G^* (Figure 2). Moreover, Table 3 (Column 8) shows, the cost value of the technical efficient inputs, resulting from multiplying the cost of each input by the technical projection input.

Table 3: Input-Output Data of 8 DMUs After Congestion Elimination

DMU	Input	Output	Cost of Input	Without congestion input cost	technical efficiency	Input projection	technical efficient input cost
	x_j^c	y_j	c_j	$C_j^* = c_j x_j^c$	θ_j^*	x_j^*	$C_j^{**} = c_j x_j^*$
A	1.5	1	2	3	1	1.5	3
B	2	2	1	2	1	2	2
C	3	3	5	15	1	3	15
D	4	4	3	12	1	4	12
E	6	5	1.5	9	1	6	9
F	7	5	1.75	12.25	0.86	6	10.5
G	7	4	3	21	0.57	4	12
H	6	3	2	12	0.50	3	6

Now, the projection inputs can be taken to the cost space $\overline{P}_c$, after being multiplied by the input costs (Table 4) (Figure 1).

Table 4: Data of 8 DMUs in Cost Form

DMU	Input $\bar{x}_j = c_j x_j^*$	Output	$\bar{\varphi}_j$
$\overline{A}$	3	1	3
$\overline{B}$	2	2	1
$\overline{C}$	15	3	1.67
$\overline{D}$	12	4	1.25
$\overline{E}$	9	5	1
$\overline{F}$	10.5	5	1
$\overline{G}$	12	4	1
$\overline{H}$	6	3	1

Now, in the cost space $\overline{P}_c$, first the cost congestion is examined by means of model (3). Table 4 (Column 4) shows the value of efficiencies. Units $\overline{B}$, $\overline{E}$, $\overline{F}$, $\overline{G}$ and $\overline{H}$ are efficient units, and input of unit $\overline{F}$ is the most one, so $x_1^{\max} = 10.5$. Therefore, inefficient units $\overline{G}$, $\overline{D}$ and $\overline{C}$ are in congestion ($15 = x_1^{\overline{C}} > x_1^*$, $12 = x_1^{\overline{D}} > x_1^*$, $12 = x_1^{\overline{G}} > x_1^*$). The decreased input of $\overline{G}$ results in $\overline{G}^c$ (Figure 2). Table 5 (Column 2) shows the value of inputs once the congestion is removed.

Table 5: Input-Output Data of 8 DMUs After Congestion Elimination

DMU	Without input cost $C_j^{****} = \bar{x}_j^c$	Output y_j	Technical efficiency $\bar{\theta}_j$	FF Input projection $\bar{x}_j^c$	Technical efficient Input cost $C_j^{*****} = \bar{x}_j^c$
$\bar{A}^c$	3	1	0.67	2	2
$\bar{B}^c$	2	2	1	2	2
$\bar{C}^c$	10.5	3	0.41	4.33	4.33
$\bar{D}^c$	10.5	4	0.63	6.67	6.67
$\bar{E}^c$	9	5	1	9	9
$\bar{F}^c$	10.5	5	0.86	9	9
$\bar{G}^c$	10.5	4	0.63	6.67	6.67
$\bar{H}^c$	6	3	0.72	4.33	4.33

Once the price congestion is removed in the cost space $\bar{P}_c$, the price technical efficiency of the units has been obtained by means of the BCC model (Table 5, Column 4). Table 5 (Column 5) shows the projection points on the frontier of $\bar{P}_c$. Among all the projection units on the frontier of $\bar{P}_c$, units $\bar{A}$ and $\bar{B}$ have the lowest cost, $C_o^{*****} = 2$. Table 6 (Column 6) shows the loss caused by the allocative inefficiency for each unit of difference is C_o^{****} and C_o^{*****} .

Table 6: Cost Decomposition

DMU	Loss due to congestion L^*	Loss due to technical inefficiency L^{**}	Loss due to price congestion L^{***}	Loss due to price tech inefficiency L^{****}	Loss due to allocative inefficiency L^{*****}
A	0	0	0	1	0
B	0	0	0	0	0
C	0	0	4.5	6.17	2.33
D	0	0	1.5	3.38	4.67
E	0	0	0	0	7
F	1	1.75	0	1.5	7
G	3	9	1.5	3.83	4.67
H	0	6	0	1.67	2.33

Table 6 shows the cost loss caused by various inefficiencies of these 8 units, and Table 7 shows the various inefficiencies.

Table 7: Decomposition of Efficiencies

DMU	Congestional eff.	technical eff.	price congestion	price tech. eff.	Allocative eff.	Overall eff. (cost eff.)
A	1	1	1	0.67	1	0.67
B	1	1	1	1	1	1
C	1	1	0.70	0.41	0.46	0.13
D	1	1	0.88	0.64	0.30	0.17
E	1	1	1	1	0.22	0.22
F	1	0.86	1	0.86	0.22	0.16
G	0.88	0.57	0.88	0.64	0.30	0.08
H	1	0.5	1	0.72	0.46	0.17

For example, cost decomposition are as follows in some units:

$$\begin{aligned} \text{Observed cost of } D &= 12 = 0 + 0 + 1.5 + 3.83 + 4.67 + 2, \\ \text{Observed cost of } F &= 12.25 = 0 + 1.75 + 0 + 1.5 + 7 + 2, \\ \text{Observed cost of } G &= 24 = 3 + 9 + 1.5 + 3.83 + 4.67 + 2. \end{aligned}$$

Unit B is cost efficient (overall efficient). The cost inefficiency of E is only due to allocative inefficiency. What follows shows the analysis of some unit's cost efficiency:

$$\text{Cost efficiency of } D = 0.028 = 1 \times 1 \times 0.88 \times 0.64 \times 0.30 \times 0.17,$$

$$\text{Cost efficiency of } G = 0.006 = 0.88 \times 0.57 \times 0.88 \times 0.64 \times 0.30 \times 0.08.$$

The proposed approach is validated with real data from twelve hospitals, examining the effects of input congestion and cost inefficiencies on overall performance. The case study confirms the method's applicability in complex, real-world settings.

5.2 Real Example

To show the applicability of this analysis, the data of 12 Hospitals, with two inputs (number of physicians and number of nurses) along with the cost of each unit and two outputs (the number of outpatients and the number of hospitalized patients) is considered, which had been used in [25] Table 8.

Table 8: Data of 12 Hospitals

DMU	Input				outputs	
	Number of doctors	Per Doctir's fee	Number of nurses	Per Nurse's fee	Number of outpatient	Number of inpatient
A	20	500	151	100	100	90
B	19	350	131	80	150	50
C	25	450	160	90	160	55
D	27	600	168	120	180	72
E	22	300	158	70	94	66
F	55	450	255	80	230	90
G	33	500	235	100	220	88
H	31	450	206	85	152	80
I	30	380	244	76	190	100
J	50	410	268	75	250	100
K	53	440	306	80	260	147
L	38	400	284	70	250	120

Table 9: Decomposition of Observed Cost into Inefficiency Components

DMU	Observed cost	Loss due to congestion	Loss due to technical inefficiency	Loss due to price congestion	Loss due to price technical inefficiency	Loss due to allocative inefficiency
A	25100	0	0	0	747.9	8802.1
B	17130	0	0	0	0	1580
C	25650	0	2998.8	0	4157.8	2943.4
D	36360	0	0	0	13691.52	7118.48
E	17660	0	2110	0	0	0
F	45150	900	5552.1	10660.74	0	12487.16
G	40000	0	0	0	12678.83	11771.17
H	31460	0	6330.4	0	3251.12	6328.48
I	29944	0	1906.84	0	0	12487.16
J	40600	0	0	0	5520	19530
K	47800	0	0	0	0	32250
L	35080	0	0	0	0	19530

Table 9 shows the cost loss caused by various efficiencies, and Table 10 shows various analyzed inefficiencies. Hospital F has two units of congestion at output 1; the congestion is a result of doctors who are present in hospital F . To remove this congestion, two doctors must be reduced, which decreases the cost by $2 \times 450 = 900$, as shown in the column related to cost loss due to congestion.

In the cost space, moreover, unit F has cost congestion at both inputs. In other words, the cost congestion and the resultant inefficiency

are due to the high salaries of doctors and nurses in Hospital *F*.

The inefficiency of Hospital *B*, which has the best performance, is only due to allocative inefficiency. In contrast, the inefficiency of the second-rate Hospital *E* is due to input value, which has caused technical inefficiency.

Table 9 shows the cost loss caused by various efficiencies, and Table 10 shows various analyzed inefficiencies. Hospital *F* has two units of congestion at the output 1; the congestion as a result of doctors who are present in hospital *F*. To remove this congestion, one has to reduce two doctors, which reduces $2 * 450 = 900$ from the costs, and Table 8 shows this in the column related to loss of costs caused by congestion. In the cost space, moreover, unit *F* has a cost congestion at both inputs. In other words, the cost congestion and the resultant are due to the high salaries of doctors and nurses in Hospital *F*. The inefficiency of Hospital *B*, which has the best performance, is only due to allocation inefficiency and the inefficiency of the second-rate *E* hospital is due to the input value that has been caused technical inefficiency.

5.3 Policy Implications

The findings of this study have several important policy and managerial implications for hospital administrators and healthcare decision-makers. First, for hospitals exhibiting congestion-driven inefficiency—such as Hospital *F*—resource reallocation should prioritize the correction of input excesses. In such cases, simply increasing the availability of personnel or resources may exacerbate inefficiency; instead, managers should focus on optimizing staffing patterns, improving scheduling systems, and aligning resource deployment with actual service demand.

Second, for hospitals characterized by technical inefficiency—such as Hospital *E*—the main policy focus should be on improving operational processes rather than reducing capacity. Measures such as workflow redesign, staff training, digitalization of administrative procedures, and better coordination across departments may help convert existing inputs into higher service outputs without requiring substantial additional investment.

Third, benchmark hospitals such as Hospital *B* should be used as reference units for best-practice transfer. Policymakers may encourage

cross-hospital learning, peer comparison, and performance monitoring systems that allow inefficient hospitals to identify operational gaps relative to frontier performers.

Finally, the cost decomposition results suggest that hospital performance should not be assessed solely on the basis of aggregate efficiency scores. Instead, policy interventions should be tailored to the specific source of inefficiency—whether congestion, technical inefficiency, or cost misallocation—so that corrective actions are both targeted and economically justified. This approach can improve not only efficiency but also the sustainability of healthcare resource allocation.

Table 10: Different Efficiencies

DMU	Congestional eff.	technical eff.	price congestion	price tech. eff.	Allocative eff.	Overall eff.
A	1	1	1	0.97	0.64	0.61
B	1	1	1	1	0.90	0.91
C	1	0.88	1	0.82	0.84	0.61
D	1	1	1	0.62	0.68	0.43
E	1	0.88	1	1	1	0.88
F	0.98	0.87	0.72	1	0.55	0.34
G	1	1	1	0.68	0.57	0.39
H	1	0.80	1	0.87	0.71	0.49
I	1	0.94	1	1	0.55	0.52
J	1	1	1	0.86	0.44	0.38
K	1	1	1	1	0.33	0.33
L	1	1	1	1	0.44	0.44

6 Conclusion

This study has developed and applied a robust framework for decomposing cost inefficiencies, providing meaningful insights into the drivers of cost performance within the healthcare sector. By utilizing a decomposition approach, this research offers a granular view of how technical, price, and allocative factors contribute to overall cost structures. Despite these contributions, several limitations must be acknowledged to contextualize the findings. First, as the empirical validation was centered on the healthcare sector, the generalizability of the results to industries with divergent cost structures requires further investigation. Second,

the precision of the decomposition is inherently linked to the accuracy of input data and price reporting. Third, while the Malmquist index allows for a dynamic perspective, the model may not fully capture sudden exogenous shocks that disrupt long-term efficiency trends. Finally, as a cost-centric framework, the model focuses primarily on resource optimization and may not explicitly account for the potential trade-offs between cost reduction and the qualitative aspects of patient care. To address these limitations and build upon the current findings, several avenues for future research are proposed. To enhance the robustness of the model against measurement errors and noise inherent in real-world data, future studies should consider integrating Stochastic Data Envelopment Analysis (SDEA) or Probabilistic DEA approaches. Furthermore, moving beyond the Malmquist index, research could investigate the velocity of congestion changes and the evolving proportions of inefficiency types to better discern whether productivity shifts stem from technological advancements or improved management strategies. Lastly, given the hierarchical nature of organizations, expanding this framework into a Network DEA model would enable a more sophisticated analysis of inter-stage interactions, exploring how efficiency improvements in one stage impact the resource requirements of subsequent stages.

References

- [1] P. L. Brockett, W. W. Cooper, Y. Wang and H. C. Shin, Inefficiency and congestion in Chinese production before and after the 1978 economic reforms, *Socio-Economic Planning Sciences*, 32(1) (1998), 1–20.
- [2] P. Byrnes and V. Valdmanis, Analyzing technical and allocative efficiency of hospitals, *Data Envelopment Analysis: Theory, Methodology, and Applications*, 129–144, Springer (1994)
- [3] A. Charnes, W. W. Cooper and E. Rhodes, Measuring the efficiency of decision making units, *European Journal of Operational Research*, 2(6) (1978), 429–444.

- [4] T. J. Coelli, D. S. P. Rao, C. J. O'Donnell and G. E. Battese, *An Introduction to Efficiency and Productivity Analysis*, Springer(2005)
- [5] W. W. Cooper, H. Deng, L. M. Seiford, and J. Zhu, Congestion: its identification and management with DEA, *Handbook on Data Envelopment Analysis*, 173–193, Springer(2011)
- [6] W. W. Cooper, L. M. Seiford and K. Tone, *Introduction to Data Envelopment Analysis and Its Uses: With DEA-Solver Software and References*, Springer.(2006)
- [7] W. W. Cooper, L. M. Seiford and J. Zhu, A unified additive model approach for evaluating inefficiency and congestion with associated measures in DEA, *Socio-Economic Planning Sciences*, 34(1) (2000), 1–25.
- [8] W. W. Cooper, R. G. Thompson and R. M. Thrall, Introduction: Extensions and new developments in DEA, *Annals of Operations Research*, 66(1) (1996), 1–45.
- [9] R. Fallahnejad, S. Asadi Rahmati and K. Moradipour, Decomposition of cost efficiency, given the set of price and cost production possibilities in data envelopment analysis, *Journal of applied research on industrial engineering*, 10.3 (2023), 392–411.
- [10] R. Fare, S. Grosskopf and C. A. K. Lovell, *Production Frontiers*, Cambridge University Press (1994)
- [11] R. Fare and S. Grosskopf, Measuring congestion in production, *Zeitschrift fur Nationalokonomie*, 43(3)(1983), 257–271.
- [12] R. Fare and S. Grosskopf, Measuring shadow price efficiency, *Applications of Modern Production Theory: Efficiency and Productivity*,(1988), 223–234,
- [13] R. Fare and S. Grosskopf, Slacks and congestion: A comment, *Socio-Economic Planning Sciences*, 34(1) (2000), 27–33.
- [14] R. Fare and S. Grosskopf, When can slacks be used to identify congestion? An answer to WW Cooper, L. Seiford and J. Zhu, *Socio-Economic Planning Sciences*, 53(3) (2001), 217–221.

- [15] R. Fare and L. Svensson, Congestion of production factors, *Econometrica*, 48(6)(1980), 1745–1753.
- [16] H.Z. Haghighi, S. Adeli, F. H. Lotfi and G.R. Jahanshahloo, Revenue congestion: An application of data envelopment analysis, *Journal of Industrial Management Optimization*, 12(4) (2016), 1311.
- [17] B. Karimi, E. Khorram and M. Moeini, Identification of congestion by means of integer-valued data envelopment analysis, *Computers industrial engineering*, 98 (2016), 513-521.
- [18] A. Khoshroo, M. Izadikhah and A. Emrouznejad, Energy efficiency and congestion considering data envelopment analysis and bounded adjusted measure: A case of tomato production, *Journal of cleaner production*, 328 (2021), 129639.
- [19] T. Kuosmanen and T. Post, Nonparametric efficiency analysis under price uncertainty: A first-order stochastic dominance approach, *Journal of Productivity Analysis*, 17(3) (2002), 183–200.
- [20] A. A. Noura, F. H. Lotfi, G. R. Jahanshahloo, S. F. Rashidi and B. R. Parker, A new method for measuring congestion in data envelopment analysis, *Socio-Economic Planning Sciences*, 44(4) (2010), 240–246.
- [21] S. Navidi, M. Rostamy-Malkhalifeh and F. H. Lotfi, Measuring congestion in data envelopment analysis without solving any models, *Scientia Iranica. Transaction E, Industrial Engineering*, 28(5) (2021), 2926-2932.
- [22] S. C. Ray and H. J. Kim, Cost efficiency in the U.S. steel industry: A nonparametric analysis using data envelopment analysis, *European Journal of Operational Research*, 80(3)(1995), 654–671.
- [23] A. Salehi, F. H. Lotfi and M. Rostamy-Malkhalifeh, A Method for identifying congestion using non-radial models in data envelopment analysis, *International journal of nonlinear analysis and applications*, 11(1)(2020), 225-235.

- [24] K.Tone, A strange case of the cost and allocative efficiencies in DEA, *Journal of the Operational Research Society*, 53(11)(2002), 1225–1231.
- [25] K. Tone and M. Tsutsui, Decomposition of cost efficiency and its application to Japanese-U.S. electric utility comparisons, *Socio-Economic Planning Sciences*, 41(2)(2007), 91–106.
- [26] H. Zare Haghighi, M. Khodabakhshi and G. R. Jahanshahloo, Review of the methods for evaluating congestion in DEA and computing output losses due to congestion, *International Journal of Industrial Mathematics*, 6(1) (2014), 1–17.

Sanaz Asadi Rahmati

Department of Basic Sciences, Technical and Vocational University (TVU),
Tehran, Iran

E-mail: S.asadirahmati@gmail.com