Journal of Productivity Analysis (2023) 60:273–294
https://doi.org/10.1007/s11123-023-00690-3

Efficiency decomposition for multi-level multi-components production technologies

Antonio Peyrache[1] · Maria C. A. Silva[2]

Accepted: 10 July 2023 / Published online: 10 August 2023

Abstract
This paper addresses the efficiency measurement of firms composed by multiple components, and assessed at different decision levels. In particular it develops models for three levels of decision/production: the subunit (production division/process), the DMU (firm) and the industry (system). For each level, inefficiency is measured using a directional distance function and the developed measures are contrasted with existing radial models. The paper also investigates how the efficiency scores computed at different levels are related to each other by proposing a decomposition into exhaustive and mutually exclusive components. The proposed method is illustrated using data on Portuguese hospitals. Since most of the topics addressed in this paper are related to more general network structures, avenues for future research are proposed and discussed.

Keywords Data Envelopment Analysis · Multi-Components Technology · Directional Distance Function · Efficiency

1 Introduction

Network models are widely used in the efficiency analysis literature. Traditional efficiency models treat decision making units (DMUs) as black boxes (transforming a set of inputs into a set of outputs), while network models have been developed to reveal the internal structure of DMUs. The origins of network models can be traced back to the contributions of Kantorovich (1939), Koopmans (1951) and Johansen (1972) (see Peyrache and Silva 2022b for a historical account of these developments). These early contributions have been reignited with the development of dynamic efficiency and the contributions of Färe and Grosskopf and some of their co-authors (see e.g., Shephard and Färe 1980; Färe 1986). Initially, dynamic DEA models (see Färe 1986; Färe and Grosskopf 1996a) modelled the idea that inputs could be divided into two sets: those that are used in a specific period, and those that can be used in any period of the time spam considered (they can be inter-temporally allocated). The authors developed models that aimed at determining the optimal time allocation of the second class of inputs. Later in Färe et al. (1997) this idea of allocation of inputs across time was extended to the allocation of land for different uses. In Färe and Grosskopf (2000) the dynamic structure of DEA models was further developed to consider that outputs produced could have two destinations: leave the system, or be used in the following year's production. The part of the output that remained in the system gave rise to what is now known as the intermediate factors that provide linkages between the different stages of the network. Interestingly the previous idea of time allocatable inputs was not much explored in the subsequent literature (an exception is Färe et al. 2010), and the modelling of these linkages has been the main avenue of research by Färe and Grosskopf and colleagues as well as many of the other contributors to the field. For example, Bogetoft et al. (2009) and Färe et al. (2018) further develop the dynamics of the network models by proposing a dynamic productivity index and comparing it with the static productivity index (where intermediates were treated as normal inputs). Nemoto and Goto (1999, 2003) were among the first to extend the ideas of dynamic network DEA introduced by Färe and Grosskopf and colleagues, to the context of cost efficiency measurement, where a static and a dynamic measure of efficiency were computed, and a dynamic effect was derived from the comparison between these two measures.

By the end of the 1990's and beginning of the 2000's network DEA models started to be shown in their multiplier

✉ Maria C. A. Silva
csilva@ucp.pt

[1] School of Economics, The University of Queensland, Brisbane, QLD, Australia

[2] Universidade Católica Portuguesa, CEGE, Porto, Portugal

form rather than in the envelopment form. Beasley (1995) is probably the best example. The author was interested in computing teaching and research efficiency of universities, where some inputs/outputs were specific to the departments and other were shared. Beasley (1995) used a multiplier model for modelling the novel situation, but not identifying the developed model as network DEA. Later Cook et al. (2000) applied a similar model to bank branches and Kao (2009b) gave rise to a number of papers on network DEA models using the multiplier model.

Since the beginning of the year 2000 the number of papers in network DEA models has grown exponentially. A google scholar search in November 2022 revealed around 1200 articles with the words 'network DEA' or 'network Data envelopment analysis' in the title. From these only 6 are prior to 2000, and 766 (around 64%) have been published after 2017. This reveals that the analysis of what happens within the black box has attracted a lot of attention in recent years.

When the black box of a DMU's production process is open, one necessarily assumes that decision making occurs at different levels in that network and involves different types of decisions, usually (but not necessarily) hierarchically organised. In the efficiency literature, the idea of hierarchical structures is not new and it has been related, many times, to the issue of resource allocation (admitting that there is a centralising hierarchically superior entity that distributes resources). Some initial research on this topic includes Färe et al. (1992), who put forward industry models with inputs that can be allocatable between firms and inputs that are firm specific. Interestingly, the models proposed in this paper resemble a lot this initial paper but in a context of processes within a firm rather than firms within an industry. Other early examples include Golany et al. (1993), Golany and Tamir (1995) who put forward a model for allocation of resources, Färe et al. (1997) that adapted the existing DEA models to the situation of one fixed and allocatable input, or Cook et al. (1998) who were among the first to come out with this notion of hierarchy in efficiency measurement. Industry models (see e.g., Lozano and Villa 2004; Lozano et al. 2004) are based on this idea of hierarchy and allocation. In fact all the original papers on network DEA have very clear the idea of re-allocating resources within the DMU or across time in case of the dynamic models. It seems that reallocation of resources is the main unifying principle underlying this stream of research. This means that opening the black-box of production allows the decision maker to make better allocation decisions across processes or the various parts of the system.

Interestingly, however, recent literature on network DEA has moved into a different direction. Recent papers, mainly those using multiplier forms of the DEA models, are mainly concerned with efficiency aggregation and/or efficiency decomposition (see e.g., Kao 2016), and very few actually address reallocation issues (see for example Kao 2014 review that put forth various network models without modelling explicitly allocation issues, or Kao 2017 that mention only briefly shared input models). The central idea of our contribution in this paper is that aggregation or decomposition of efficiency measures, happening at different decision levels, needs to take into account resource allocation issues. More specifically, processes efficiencies cannot be simply aggregated to form an overall efficiency score of the unit, unless there is independence between the parts being aggregated. Whenever there is interdependence between the parts the overall efficiency score of the DMU is necessarily constituted by a part of efficiency that is due to the parts operating inefficiently and another part that is due to the misallocation of resources across the various parts of the whole (by the DMU's decision maker). These issues are addressed in this paper for a specific type of network model: parallel network structures (see e.g., Kao and Hwang 2010, or Castelli et al. 2010 for classifications of network structures), but the rationale is in fact extended to all types of network structures. In particular a number of models are proposed to measure inefficiency at different levels of decision, and to decompose higher level's inefficiency into the lower level inefficiency and reallocation inefficiency. In parallel systems this decomposition, to the authors' knowledge, has not been proposed before.

A parallel or multi-component structure of processes within a DMU can be encountered in various situations. The DMU can be seen as a firm or a production unit. For example, a hospital whose manager is facing the problem of allocating efficiently the resources she is given among the different departments of the hospital (i.e., cardiology, radiology, etc.) and leaves the decision on how to best use these resources to the department head. The DMU can alternatively be an entire industry or sector. For example, a central planner in a government agency facing the problem of efficiently allocating resources (factors of production or expenditure) to individual hospitals, courts of justice, schools, etc. In doing so, the central planner decides the quantity of resources that go to each single hospital, court or school and then leaves the decision on how to best use these resources to the manager. As a result, parallel structures are intrinsically linked with industry models (or centralised allocation models), but this link has not been much noticed. In addition and as a result, parallel network structures are also intrinsically linked to the problem of the efficient allocation of scarce resources between processes and multilevel decision models (where allocation decisions are performed at various levels).

In this paper, we propose some models for assessing the efficiency of each hierarchical level in a coherent way. Note

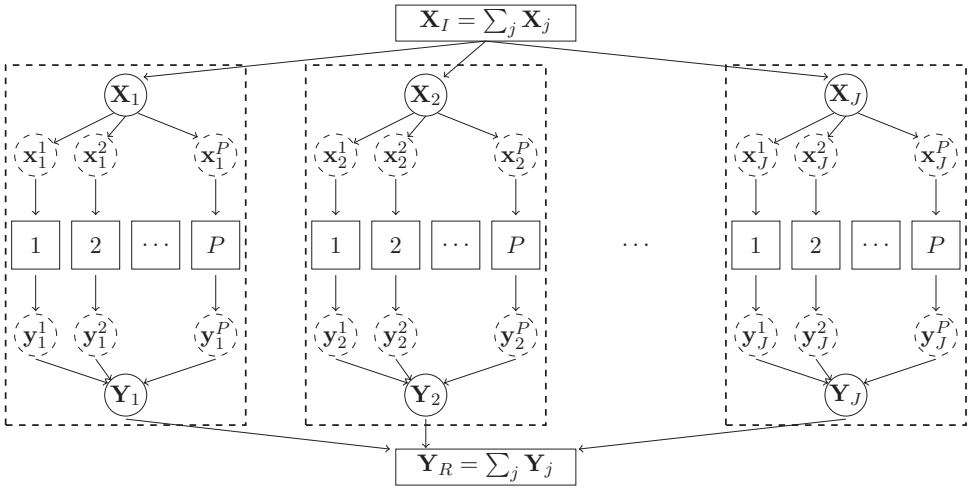


Fig. 1 Graphical representation of a parallel network system

that production is carried out at the process level, and resource allocation decisions are made at the firm and industry levels. As a consequence of this fact, in our efficiency analysis we distinguish between inefficiencies arising from non-optimal allocation of resources and inefficiencies arising from misuse of production factors during production. How we define, relate and attribute these different types of inefficiency to the different levels is the core of our contribution. Another contribution includes the reconciliation of previous existing network models in the literature, where we call the attention for the fact that many of these are complementary rather than alternative, as they look at different levels of assessment.

The paper is organised in the following way. In Section 2, we discuss multi-component structures, present two models and compare them using an illustrative numerical example. In Section 3, we introduce a classification of the different types of inputs and outputs and their associated data matrices. In Section 4, we propose our approach to the measurement of efficiency for the network as a whole. In Section 5, we apply our own approach to the numerical illustrative example. In Section 6, we provide an empirical application to Portuguese hospitals. In Section 7, we conclude.

2 Multi-component structures

2.1 The setting

Consider an industry (or system or network) composed of a group of firms (or DMUs) $j = 1, \ldots, J$ and the production process components (or sub-DMUs) within each firm $p = 1, \ldots, P$. Production processes use $i = 1, \ldots, I$ inputs (factors of production) to produce $r = 1, \ldots, R$ outputs (final products) (different processes can use different subsets of the inputs and produce different subsets of outputs). The quantity of input i of subunit p in unit j is denoted by x_{ij}^p, and the quantity of output r of subunit p in unit j is denoted by y_{rj}^p. The overall quantity of input i available to DMU j is indicated with a capital letter and is equal to the sum across processes: $X_{ij} = \sum_{p=1}^{P} x_{ij}^p$. Similarly, the overall quantity of output r produced by DMU j is $Y_{rj} = \sum_{p=1}^{P} y_{rj}^p$. In a black box approach one would use these overall quantities to assess the efficiency of the DMU. Parallel network structures can be depicted as in Fig. 1, where a DMU is composed of P subunits or processes, using a number of similar inputs to produce a number of similar outputs. The overall quantity of input available to the industry is $\mathbf{X}_I$. Each individual firm is allocated a quantity $\mathbf{X}_j$ and within the firm this quantity is allocated to the various production processes to produce the final quantity of output $\mathbf{Y}_j$. Summing these production quantities across firms returns the overall quantity of output produced by the industry $\mathbf{Y}_R$.

2.2 The literature

In order to clarify our understanding of what is a parallel network structure we start with the classification of Castelli et al. (2010). These authors use the term *Elementary units* to refer to those units whose subunits: (i) do not share inputs, in the sense that the DMU cannot make decisions on how to allocate resources; (ii) have similar inputs and outputs, and all inputs and outputs of the DMU are also inputs of its subunits; and (iii) are not linked by flows of intermediate materials. Elementary units are therefore the simplest case, where all subunits are independent and therefore the main problem regarding efficiency measurement is the computation of each subunit efficiency (which is independent from the rest) and then to aggregate these efficiencies, if relevant,

in a manner that allows one to reflect the efficiency of the whole DMU. Naming 'network' this elementary structure of independent subunits is somehow misleading, since a network implies some sort of interdependence and interconnection. In a parallel network system this interdependence exists in the form of shared resources that the DMU distributes (allocates) to its subunits. This means that assumptions (ii) and (iii) are in general met in a parallel network production model (but assumption (i) is not).

Before presenting the models that have been used in the literature to handle parallel network models, it is critical to stress that when subunits are independent (i.e., the DMU is elementary), the DMU efficiency is the aggregate of its subunit efficiencies. This is in fact the definition of structural efficiency introduced by Farrell (1957), where the author stated that structural efficiency was the weighted average of the efficiency of the constituent firms of an industry. The approach to follow in the case of independence between processes is therefore a bottom-up approach, where one first assesses the efficiency of constituent firms (in an industry setting) or processes (in a firm setting), and then aggregates these efficiencies to obtain an industry or a DMU overall efficiency score.

When subunits are not independent we assume implicitly that there is a central decision maker who allocates resources to each of the processes or to each of the DMUs in an industry. Under this assumption the sum of all inputs (or a subset of them) of the processes into a DMU input vector, and the sum of all outputs (or a subset of them) of the processes into a DMU output vector makes sense since by summing inputs and outputs one implicitly assumes that there is a total amount of inputs at the disposal of the DMU that can be used to produce a total amount of outputs (note that if processes are independent, this sum does not make sense, since inputs are process specific and only relevant at the subunit level).

Industry models assess the efficiency of an average production unit and take this efficiency as the aggregate efficiency of the industry (see e.g., Asmild et al. 2009). The debate on how the performance of the average DMU differs from the aggregate performance of all firms, started in Ylvinger (2000) that advocates that the use of the average unit (as initiated in a study of Forsund and Hjalmarsson 1979) cannot measure the efficiency of the industry. Later on Li and Cheng (2007) showed that structural efficiency and the efficiency of the average unit are equivalent concepts under an identical convex individual technology set, and that differences between the two are related to allocative efficiency. Karagiannis (2015) explored in more depth the relationship between the efficiency of the average unit and structural efficiency. The author concludes that the two concepts of efficiency will coincide only if size is uncorrelated with efficiency and if there are no reallocation inefficiencies. The efficiency of the average DMU has been explored by several authors under the denomination of centralised allocation models or industry models (e.g., Lozano and Villa 2004; Peyrache 2013; Peyrache 2015). Note that Färe et al. (1992) is indeed one of the first papers mentioning industry performance and relate that with firm performance.

If we regard firms in industry models as the basic processes, and the industry as the DMU, then the aggregation problem as set out in the industry models can be used to assess the efficiency of the parallel network structure. The main difference between an industry model and a parallel network model is that in the latter case the industry is a third level of decision making on top of the firm and the process. In other words, in a parallel production network we have processes $->$ firms $->$ industry; rather than just firms $->$ industry as in typical industry model.

Linked with the industry efficiency is also another strand of the literature: that of developing common set of weights to assess DMUs. This strand has very different objectives from the above—usually the selection of the best unit is the purpose, while other times it is to ensure fair comparisons between units - but the mathematical formulations end up being very similar to industry models or centralised resource allocation models. See Afsharian et al. (2021) for a recent review on this topic and links with other strands of the literature.

In the case of independence between processes, all one needs to decide is on an aggregation rule for the individual efficiencies of the subunits that may yield a satisfactory aggregate measure (see e.g Portela et al. 2016). When processes are not independent then reallocations are possible between processes and as a result not only technical, but also reallocation efficiencies are of interest. The literature on parallel network structures has followed mainly the former approach (see. e.g., Kao 2009b; Kao 2009a; Kao 2012), therefore disregarding re-allocation efficiencies. The related literature on series network models or relational models has also been very much concerned with the aggregation of efficiency scores on process to obtain an efficiency score at the DMU level disregarding re-allocation efficiencies (see e.g., Kao 2018, Kao 2016). Between the two extreme assumptions pointed out above (complete independence between processes and complete possibility of resource reallocation) there may be several other possible situations: for example some inputs and outputs can be allocatable across units, but others may be process specific and non-allocatable. This situation is related to the literature on output specific inputs modelling, where the production of each output is modelled through a different production possibility set (see e.g., Banker 1992 who firstly introduced the concept of separable production functions and Cherchye et al. 2013 who applied it). Output-specific inputs can be modelled together

with joint inputs or shared inputs, which are those that are shared or that can be allocated to various processes. In Cherchye et al. (2013) this type of inputs are called joint inputs, and in Castelli et al. (2010) this type of models are called 'Shared flow models'. One of the first application of Shared flow models was by Beasley (1995) (see also Mar Molinero 1996) in an analysis of university departments where teaching and research were considered two separate functions consuming some joint inputs (e.g., equipment expenditure). Cook and Green (2004) also applied this type of models but called them 'multicomponent model structures'. In these models the proportion of shared input that is allocated to each department is taken as a decision variable (and therefore the allocation is non-observable).

Another structure considered in Castelli et al. (2010) is that of multilevel models - where there may be inputs that are used by the DMU but not by any of its subunits. Cook et al. (1998) called these models hierarchical models and argue that traditional models cannot be applied to DMUs that are somehow grouped in a hierarchical form, since factors that are produced or used at one level (e.g., the hospital level) may not be produced or used at another level (e.g., the hospital service level). They developed models in two stages, where units are assessed within groups and also at a higher hierarchical level. The way the authors link the two assessments is by adjusting the within group scores by a factor that takes into account the higher level efficiency scores. The basic idea of Cook et al. (1998), is in a sense related to what we do in this paper, but the way of implementation and analysis is completely different.

Summing up, there is much literature related to the parallel network structures but the links have not been recognised. In addition the proliferation of different names for similar models, or different names for special types of inputs and/or outputs has not helped in terms of making the required links in the otherwise sparse literature. We believe this paper contributes to clarifying some of these issues and reconcile the previous literature on similar/related topics.

2.3 Existing parallel network models

Seminal papers on Network DEA considered two stage networks (see Färe and Grosskopf 1996b) and dynamic models (see Färe and Grosskopf 2000), both including the existence of intermediate flow variables.

If each subunit is a representation of the same DMU observed in different periods of time, we would fall in the realm of dynamic efficiency models. The main difference would be that in dynamic models there are some intermediate variables that flow from one period to the next, while in parallel models there are no intermediates.

If we ignore intermediates, the structure of parallel network models is the same as the one of dynamic models. Therefore we can use the models developed in the dynamic setting without intermediates to set the scene for two technologies that have been used in modelling parallel network models: the process specific technology and the system technology. See the explicit formulation of these technologies in Lozano (2011) who also adopted this distinction.

Following Färe and Grosskopf (2000) the network dynamic model without intermediates and an input orientation is shown in (1) (see also Färe et al. 2007).

$$\begin{aligned} &\min_{\lambda_j^p,\theta^o} \theta^o \\ &st \sum_{j=1}^{J} \lambda_j^p x_{ij}^p \le \theta^o x_{io}^p, \quad \forall i,p \\ &\sum_{j=1}^{J} \lambda_j^p y_{rj}^p \ge y_{ro}^p, \quad \forall r,p \\ &\lambda_j^p \ge 0, \quad \forall j,p \end{aligned} \tag{1}$$

This model sets a technology for each subunit and models it through p constraints that are subunit or process specific. This modelling is related to recent developments on output-specific inputs. In the output-specific input literature one assumes that not all inputs have an impact on the production of all outputs, and, as a result, different technologies are associated with different sets of inputs and outputs. Cherchye et al. (2013) and Cherchye et al. (2017) propose models that can handle process specific and shared inputs (or 'joint inputs' as they named them). These models assume that joint inputs are simultaneously used by all processes and cannot be distributed or allocated between processes (behaving therefore as a public good). The recent literature on output-specific inputs is also related to earlier literature addressing the topic of separable technologies as in Banker (1992), or to literature that addressed the issue of disaggregating the traditional DEA ratio of aggregated outputs to aggregated inputs into partial ratios of outputs to inputs and modelling these individual ratios into a general model as in Salerian and Chan (2005), Despić et al. (2007), or Silva (2018)). Recently Podinovski et al. (2018) propose a multiple hybrid returns to scale (MHRS) technology, and they assume that shared inputs (or outputs) may be perfectly joint in the sense of Cherchye et al. (2013) or fully allocated, although in unknown proportions.

An alternative model for dynamic structures is due to Kao (2013) and is based on a technology defined as the aggregate of all processes—which we call system technology. Such model is shown in (2) for the situation where one ignores constraints on intermediates, and assumes an input

orientation.

$$\begin{aligned} &\min_{\lambda_j^p,\theta^o} \theta^o \\ &st \sum_{p=1}^{P}\sum_{j=1}^{J} \lambda_j^p x_{ij}^p \le \theta^o X_{io}, \quad \forall i \\ &\sum_{p=1}^{P}\sum_{j=1}^{J} \lambda_j^p y_{rj}^p \ge Y_{ro}, \quad \forall r \end{aligned} \tag{2}$$

where X_{io} is the total sum of input i available to firm o and Y_{ro} is the total sum of output r produced by firm o. Model (2) is presented in Kao (2009a, 2012, 2013) as the parallel network model. The model is however usually shown in the multiplier form, contrarily to the original literature on Network DEA models that used the envelopment form.

The multiplier model of Kao (2012) is shown in (3), where u_r is the weight assigned to output r and v_i is the input weight assigned to input i - weights are considered the same across subunits (i.e., the implicit value attributed to each input and output should be the same in each subunit). Note that the original model has more constraints, but some are redundant. As a result, we simplified the model of Kao (2012) by excluding redundant constraints and ignoring slacks. This results in model (3) being the dual of model (2).

$$\begin{aligned} &\max_{u_r,v_i} E_o = \sum_{r=1}^{R} u_r Y_{ro} \\ &st \quad \sum_{r=1}^{R} u_r y_{rj}^p - \sum_{i=1}^{I} v_i x_{ij}^p \le 0, \forall j, \quad \forall p \\ &\sum_{i=1}^{I} v_i X_{io} = 1, \\ &u_r, v_i \ge 0 \end{aligned} \tag{3}$$

According to Kao (2012), model (3) results in efficiency scores for each DMU_o (E_o^*). The efficiency of subunit p in DMU j (e_j^p) is determined using\mathop{\sum }\limits_{i=1}^{I}{v}_{i}{X}_{io}=1,\\ {u}_{r},{v}_{i}\ge 0 the optimal weights of model (3) identified with an * in (4):

$$e_j^p = \frac{\sum_{r=1}^{R} u_r^* y_{rj}^p}{\sum_{i=1}^{I} v_i^* x_{ij}^p} \tag{4}$$

The computation of subunits efficiency in this way allows that the DMU efficiency can be decomposed into the efficiency of the subunits using (5) (based on Kao 2012):

$$E_j^* = \sum_{p=1}^{P} w^p e_j^p, \quad \text{where} \quad w^p = \frac{\sum_{i=1}^{I} v_i^* x_{ij}^p}{\sum_{i=1}^{I} v_i^* X_{ij}} \tag{5}$$

Lozano and Villa (2004) have also proposed similar models to Eqs. (2) and (3) in the context of assessing industry efficiency (called centralised radial resource allocation models) where each subunit p in the above models corresponds to a DMU. The authors note (p. 149) that "the objective function represents the efficiency of an aggregate unit representing the average of all existing DMUs", and, as we saw before, is taken as the aggregate industry efficiency, which differs from structural efficiency. Kuosmanen et al. (2006)) also proposed similar models for analysing the industry cost efficiency.

The above models have been proposed in the literature under CRS, without any clear economic reason for this restrictive assumption on returns to scale. Under VRS, a set of P constraints is added to model (2) stating that $\sum_{j=1}^{J} \lambda_j^p = 1, \forall p = 1, \ldots, P$. This corresponds, in the multiplier model (3), to the addition of P free variables u_o^p in the objective function and in the first set of constraints. This implies that the efficiency of subunit p under VRS is given by:

$$e_j^p = \frac{\sum_{r=1}^{R} u_r^* y_{rj}^p + u_o^p}{\sum_{i=1}^{I} v_i^* x_{ij}^p}.$$

In the system technology the same inputs and the same outputs are linearly combined between processes and compared to an aggregate which is the sum of the inputs and the sum of the outputs for the DMU. This results in the efficiency score obtained for the DMU being in fact a score of an 'average' process (See for a discussion and proof Lozano and Villa 2004 p.149). Under CRS the technology against which this average DMU is assessed is the technology of the most productive process, and under VRS the technology is not the intersection of processes technologies but is enlarged by all possible combinations of processes. For example in model (2) the aggregate input i of unit o being assessed can be compared with a combination of the input i from p different processes.

2.4 Illustrating and comparing the two approaches

We use an illustrative example with 4 DMUs each composed of 3 subunits, each using a single resource to produce two outputs. Table 1 shows the data for this example.

Using input oriented models, we will consider 4 possibilities for assessment: (i) assess the efficiency of the processes of the 4 DMUs independently using the standard model of Charnes et al. (1978); (ii) assess the efficiency of the DMU ignoring the processes (black box approach), (iii) use model (1) to obtain the efficiency of the DMU; (iv) use model (3) to obtain the efficiency of the DMU and the processes simultaneously.

Table 2 shows process efficiencies computed independently, the aggregate of these efficiencies (where the aggregation weights were determined by the share of the input used) and the black box efficiency.

Table 1 Data for illustrative example

DMU	X_j	Y_1j	Y_2j	PROC 1			PROC 2			PROC 3		
				x_j^1	y_{1j}^1	y_{2j}^1	x_j^2	y_{1j}^2	y_{2j}^2	x_j^3	y_{1j}^3	y_{2j}^3
1	120	75	100	30	40	60	60	25	20	30	10	20
2	100	57	85	40	25	20	40	22	25	20	10	40
3	130	84	215	40	30	65	30	14	20	60	40	130
4	260	128	170	45	30	60	200	90	100	15	8	10

Table 2 Subunit efficiencies evaluated separately and aggregated at the DMU level

DMU	Proc1	Proc2	Proc3	Aggregate	Black Box
	CRS				
DMU1	100.0%	75.76%	50.00%	75.38%	96.73%
DMU2	46.87%	100.00%	92.31%	77.21%	88.21%
DMU3	81.25%	100.00%	100.00%	94.23%	100.00%
DMU4	66.67%	81.81%	80.00%	79.09%	76.19%
	VRS				
DMU1	100.0%	78.43%	60.19%	79.17%	100.00%
DMU2	75.00%	100.00%	100.00%	77.21%	100.00%
DMU3	100.00%	100.00%	100.00%	100.00%	100.00%
DMU4	66.67%	100.00%	100.00%	94.23%	100.00%

Under CRS when one evaluates the DMUs ignoring their subunits we have that DMU 3 is the only efficient DMU. DMUs 1, 2, and 4 are inefficient. A separate evaluation of the subunits yields that DMU1 is the one showing efficiency in process 1. Process 2, on the contrary, is efficient in DMUs 2 and 3, while process 3 is efficient in DMU3. All DMUs have at least one subunit efficient, except DMU4 where all its subunits are inefficient. Under VRS, the black box approach yields all DMUs efficient, but this result is only consistent with processes efficiencies in the case of DMU 3 that shows all processes 100% efficient. All the remaining DMUs have at least one process that is inefficient under VRS.

When one applies models (1) and (2) the results are as shown in Table 3.

As shown before, results from model (1) equal the maximum process efficiencies (when assessed independently). Results from model (2) may result in all DMUs being inefficient as is the case of CRS. In this case, DMU 3 is the most efficient unit and DMU 4 is the least efficient unit, both under CRS and VRS. For the CRS case, these results are consistent with the evaluation of the DMUs disregarding their internal structure (the black box approach) shown in the last column of Table 2.

Models (1) and (2) do not allow the computation of process efficiency scores (see also Chen et al. 2013). For

Table 3 Efficiencies of DMUs

DMU	Model (1) CRS	Model (2) CRS	Model (1) VRS	Model (2) VRS
DMU1	100%	46.88%	100%	76.70%
DMU2	100%	42.75%	100%	75.00%
DMU3	100%	77.80%	100%	100%
DMU4	81.82%	36.92%	100%	73.53%

Table 4 Subunit efficiencies under model (3)

DMU	$Proc_1$	$Proc_2$	$Proc_3$
DMU1	100.00%	31.25%	25.00%
DMU2	46.88%	41.25%	37.50%
DMU3	77.68%	33.57%	100.00%
DMU4	50.01%	33.76%	40.01%

the case of the system technology one can use the multiplier model and follow the procedure in (4) and (5) which allows the computation of process efficiencies through the optimal weights obtained from solving model (3). The efficiency of the subunits are shown in Table 4 for the case of CRS.

In the VRS case inconsistent results, like negative efficiency scores, are found from the application of the optimal weights (for details see Peyrache and Silva 2022a). As a result, the procedure of applying optimal weights obtained from model (3) to estimate process efficiencies is not well defined under VRS.

Note also that some counter-intuitive results for the efficiency of the subunits are obtained under the CRS model (3). In particular, process 2 shows up as highly inefficient for all DMUs. However, when assessed individually this process in fact shows the highest efficiency scores with two DMUs being efficient in this process. The main problem from using optimal weights from model (3) to assess process efficiencies is related to their non-uniqueness. Kuosmanen et al. (2006) also mentions the problem of non-unique weights in an industry model, noting that it results in a problem just for subunits efficiency and not for the DMUs efficiency, meaning that these aggregate models are not well fit to obtain subunits' efficiency but just for assessing the DMU's efficiency.

In addition to the above, the multiplier models cannot provide targets for DMU's and processes. One needs to use envelopment models to get this information. However, envelopment models yield targets that are inconsistent with the efficiency scores based on the weights. This can be seen in Table 5, where the lambda values for all DMUS are 0 under CRS yielding therefore null targets for thisprocess in all DMUs.

Table 5 Envelopment Results for model (2) under CRS and VRS

	CRS				VRS			
	DMU1	DMU2	DMU3	DMU4	DMU1	DMU2	DMU3	DMU4
Eff	0.4688	0.4275	0.7780	0.3692	0.767	0.75	1	0.7353
λ_1^1	1.875	1.425	0.829	3.2	1	1	1	1
λ_2^1	0	0	0	0	0	0	0	0
λ_3^1	0	0	0	0	0	0	0	0
λ_4^1	0	0	0	0	0	0	0	0
λ_1^2	0	0	0	0	0	0	0	0
λ_2^2	0	0	0	0	1	0	1	0.6177
λ_3^2	0	0	0	0	0	1	0	0
λ_4^2	0	0	0	0	0	0	0	0.3824
λ_1^3	0	0	0	0	0	0	0	0
λ_2^3	0	0	0	0	0	0	0	0
λ_3^3	0	0	1.271	0	0.15625	0	1	1
λ_4^3	0	0	0	0	0.84375	1	0	0

This inconsistency (of having subunits efficiency scores different from zero and target levels of zero) has been noted elsewhere and therefore we will not detail on the issue. For example, Pachkova (2009) addressed this issue of targets being zero, interpreting it as the indication for closing down Sub-process. This implies that in models such as (2) under CRS processes are considered completely reallocatable and substitute. Chen et al. (2013) also mentions that "the divisional efficiency scores obtained from the multiplier model can be unfeasible under the envelopment model under the condition of CRS" and suggest that envelopment models should not be used to find divisional efficiency scores (see also Lim and Zhu 2016 that develop further on this issue).

One way to sort out the above problem can be the use of a VRS model, which forces all subunits to remain active for each DMU through the convexity constraint imposed for each subunit. This, however, only sorts the problem of avoiding the closure of subunits, since the inconsistency between targets and subunits efficiency scores obtained from (4) remains.

The technology specific model (1) shows a similar problem: It is able to provide targets for each subunit both under CRS and VRS, but there is no immediate way to obtain subunit's efficiency scores.

2.5 Summary of the section

Summing up, under model (1) targets can be computed and are similar to targets obtained in independent assessments, and in model (2) targets may also be computed, but they are in general inconsistent with process efficiencies, because there may be targets of zero and efficiency scores different from zero. Therefore model (2), when analysed from the target perspective, may suggest the closure of inefficient processes and the maintenance of the most productive processes only (which implies perfect substitution between processes). This issue has also been identified by Pachkova (2009) on full reallocation models. Full reallocation is in fact the allocation assumption implicit in model (2). On the contrary process-specific technologies, as those applied in output-specific input settings, in general yield the efficiency of the DMU as being the same as the maximum efficiency across its processes (and therefore, disregards completely inefficient processes).

As a result, the literature has not reached a consensus on the type of technology that is more appropriate to handle a network parallel model, nor does it have a framework that is able to provide simultaneously the efficiency of the DMU and of its processes in a meaningful and coherent way.

Our view in this paper is that the above problems can be solved by making explicit the assumptions that underlay the construction of these models, and by recognising that the aggregate efficiency of the DMU is not the sum of the efficiency of its parts. For example, the system technology model of (2) implicitly assumes that inputs and outputs from different processes can be fully allocated and are not process specific. On the contrary, the process specific technology considers the opposite - that inputs and outputs are process specific and cannot be allocated. Under the first assumption it is reasonable to decide to close down some processes leaving just those that are most productive. Under the second assumption processes are treated as completely independent and the DMU efficiency that model (1) returns is in fact not an aggregate of efficiencies, but its maximum. In real situations what we may have is something between these two

extremes. This implies recognising that one may have inputs and outputs that are subunit specific (and therefore cannot be aggregated and cannot be re-allocated); and we may have allocatable inputs and outputs, whose allocation is known, and have been allocated to processes by a central decision maker, and we may also have public inputs and outputs that are not allocated to any process but can be used or produced by all of them (that is, the use of one resource by one department does not prevent others from using the same resource).

The clear definition of the type of inputs and outputs used in the model is very important for the whole analysis. In this paper, we are going to distinguish between:

- (i) inputs and outputs that are process specific;
- (ii) inputs and outputs that are allocatable, and the allocation is observed;
- (iii) inputs and outputs that are public goods or joint non-allocatable (they can be used by one process without preventing use by another process);
- (iv) those that are allocatable but the allocation is non-observed.

The distinction between different types of factors is not new in literature, but authors have referred to the same class with different names. For example, the term joint inputs or shared inputs has been used to consider those factors whose use is shared by all processes but the allocation is not observed (iii). In Beasley (1995) (see also Mar Molinero 1996) it was assumed that although not observed this allocation could be determined. Similarly Cherchye et al. (2013) also distinguished between joint inputs and output specific inputs, but they argued that allocation is not to be determined because shared outputs are jointly used by all processes and its total amount is available to all (see also Cherchye et al. 2015) who introduced the concept of sub-joint inputs and adapted their approach to the case of undesirable outputs. Podinovski et al. (2018) (and also more recently Podinovski 2022) use the term shared inputs and assume that they can vary between the two extremes of perfectly joint (behaving as public goods) and fully allocated in unknown proportions. Our denomination of joint non-allocatable inputs is therefore more in line with the concept of Cherchye et al. (2013), since we assume that when the joint inputs are fully allocatable the allocation is known and therefore we call them allocatable inputs. Note that our denomination of joint non-allocatable inputs can also incorporate another category of inputs and outputs: those that are not proper variables at the process level and are important variables at the firm level. As a result emergency costs in an hospital or emergency patients can be considered a joint input and a joint output, since they are not observed at the level of the service or specialty but only at the level of the hospital. There are not many options in the literature to handle this multi-level data in DEA. To the best of our knowledge the only study that indeed considered this multi-level structure of data into DEA and treated it as a network structure was that of Cook et al. (1998)—some other models that are deemed multi-level or hierarchical indeed do not fall into this category. For example, Cook and Green (2005) considered that the DMU level variables are allocatable in unobserved proportions to the processes and the model fall in the Beasley (1995) type models.

We should take notice at this point of the fact that category (iii) and (iv) not only are used sometimes interchangeably, but they are providing alternative ways of solving the same problem. In practical terms, if an input is allocatable but the allocation is not known, it could be treated as a joint input. In purely methodological terms this is incorrect and one should seek to collect more data on the allocation to the different subunits. How to treat allocatable inputs whose allocation is not observed is still an open issue in the literature. Podinovski et al. (2018) provide a solution for the CRS (scalable) technology. Extensions of their ideas to the VRS and non-convex case would be an important avenue of research. In fact, one may argue that there are very few public goods, and in most cases, in practice, the input is allocatable without observing its allocation. Although this is basically a problem of lack of data, some methodological advances are still possible in this framework.

One contribution of this paper is to show that the two technologies presented before (the process-specific technology and the system technology) are not alternatives but complements, as they allow the estimation of efficiency at different hierarchical levels. Having these definitions in mind, the next sections will propose models that allow one to solve the problems that we raised in this section.

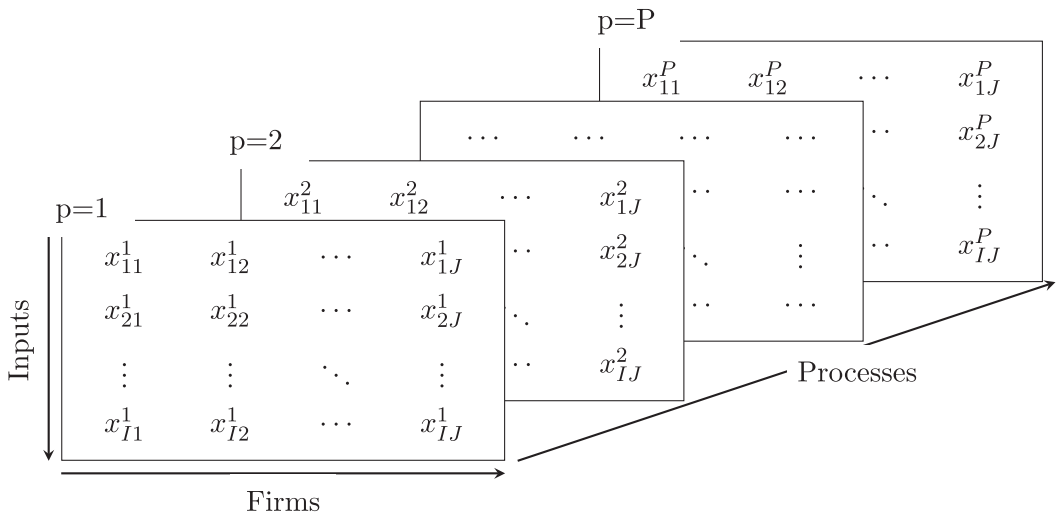
3 The data matrices: allocatable, process specific and joint inputs and outputs

In this section, we show that by building the data matrices for the inputs and the outputs carefully, one can include the three forms of inputs and outputs (allocatable, process specific and joint) in a standard, convenient and parsimonious notation. This will also help in clarifying how to include these different types of inputs and outputs in the model. In order to distinguish between the different types of inputs and outputs, we consider the *observed* allocation of

inputs for a given process p (this is the actual dataset at hand):

$$\begin{bmatrix} x_{11}^p & x_{12}^p & \dots & x_{1J}^p \\ x_{21}^p & x_{22}^p & \dots & x_{2J}^p \\ \vdots & \vdots & \ddots & \vdots \\ x_{I1}^p & x_{I2}^p & \dots & x_{IJ}^p \end{bmatrix}, \quad \forall p \tag{6}$$

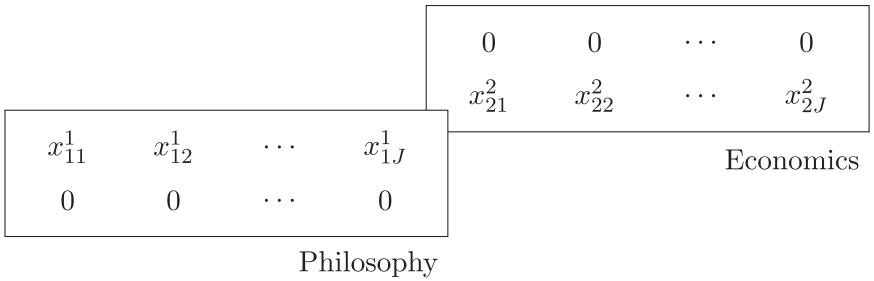
Since there is one matrix of data for each process p we can stack them together to have a visualisation of the whole dataset



where the generic element total available input for firm j is the raw sum $\sum_p x_{ij}^p$ and the total input available at the system (or industry) level is $\sum_j \sum_p x_{ij}^p$. This can be collected into a $I \times 1$ vector ($\mathbf{x}_j^p$) representing the use of all inputs for process p in firm j (this is equivalent at looking at a particular row of one of these matrices). We say that input i is allocatable if it can be freely (at no cost) reallocated across processes. As an example of an allocatable input, one could think of beds in a hospital: these can be reallocated across the different specialties (or processes of the hospital) at negligible cost. We say that a factor of production i is perfectly allocatable if it can be allocated to any of the processes. Since the allocation possibilities for input i are described by the rows of the data matrix, perfect allocatability requires that all the coefficients in the associated row are positive for at least one firm $\sum_j x_{ij}^p > 0, \forall p$. If an input is only allocatable to a subset of the processes, we say that it is partially allocatable and some of the associated coefficients in the matrix will be equal to zero. In particular, if input i cannot be used in process p, then we should observe that $\sum_j x_{ij}^p = 0$. Finally, we say that an input is process specific if it can only be allocated and used by a specific process in the network. These are very generic requirements, in the sense that they are necessary but not sufficient to the classification of inputs and outputs as allocatable or not. For example, if we observe that all the entries in all firms for a particular input in a particular process are equal to zero, then this is sufficient to state that the input is not allocatable to that process. In a more formal way, if $\sum_j x_{ij}^p = 0$ then input i is not used in process p. On the contrary, the fact that we observe one input allocated to one particular process does not imply that the input can be reallocated at no cost. In other words, although the researcher will look at the dataset to infer if an input is allocatable, this will be far from sufficient to establish if it falls in any of the previous categories. To say this in yet another way, the researcher will have to decide ex-ante if an input is allocatable or process specific or joint; and she should define a classification of variables that does not contradict the basic data requirements discussed above.

As an example, suppose that the first line of each matrix p is the number of teachers used in each university and suppose that there are only two departments: Economics and Philosophy. Since the teachers in economics cannot be freely (at no cost) reallocated to the philosophy department (and vice-versa), the way one needs to represent this is by adding a second row. Therefore from one input (number of teachers), one artificially builds two inputs: the number of teachers in economics and the number of teachers in philosophy. The entries for the number of philosophers in the economics department will all be equal to zero (and vice-versa). Therefore when one is summing up the inputs to the university level, the overall number of economists and philosophers will be the same, i.e., no reallocation of

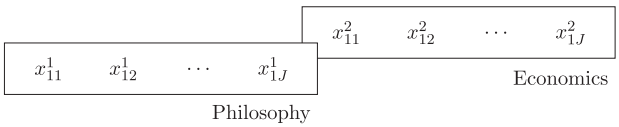
teachers is possible. For the sake of the example, the previous data matrix will be:



If we take the sum of inputs across processes, then we obtain a $J \times 2$ matrix (since there are two inputs):

$$\begin{bmatrix} x^1_{11} & x^1_{12} & \dots & x^1_{1J} \\ x^2_{21} & x^2_{22} & \dots & x^2_{2J} \end{bmatrix} \tag{7}$$

Thinking of the same example, if teachers were allocatable across the two departments, then the data matrix would be build as:



and the sum of the only allocatable input (the number of teachers) would be the following vector:

$$\left[\left(x^1_{11} + x^2_{11}\right) \quad \left(x^1_{12} + x^2_{12}\right) \quad \dots \quad \left(x^1_{1J} + x^2_{1J}\right) \right] \tag{8}$$

According to this second classification of the input, each university can choose how to allocate it across the two processes.

The inputs considered so far are observed and allocated at the process level. Conceptually, there is another distinct group of inputs that are only observed at the firm level and we called in the previous section joint inputs. These inputs are not allocated to any specific process within the firm and are available in the same quantity to all processes, which means they have a public good nature. An input (similarly for an output) is said to be public or shared if it is non-rivalrous in production. This means that if a certain quantity is provided to process p, this same quantity can be used at no cost on all the other active connections of the system. If we call the total amount of public input i available to the firm x_{ij}, then this quantity of public good will be the same for all processes of the firm $x^p_{ij} = x_{ij}$ itself.

One should be aware that our classification of inputs and outputs is known beforehand and is not in any way inferred using our model. In this sense the horizon and the scope of the analysis will determine what is allocatable and what is specific to the process. It may happen, for example, that a specific resource is process specific in the short run but allocatable in the long run. The previous discussion basically reduces our notation to the standard case, with the caveat that we have to make sure all programmes are feasible if the data (both inputs and outputs) contains zeros.

4 The inefficiency of the parallel network model

We start with the inefficiency of each single process by considering model (9), where we solve for all processes of firm o in one single step. Note that the input and output matrices and vectors in the next models are as defined above and can include various types of inputs and outputs.

$$\begin{aligned} &\max_{\beta^p, \lambda^p_j} \sum_p \beta^p_o \\ &st \sum_{j=1}^{J} \lambda^p_j \mathbf{x}^p_j \le \mathbf{x}^p_o - \beta^p_o \mathbf{g} \ \forall p \\ &\sum_{j=1}^{J} \lambda^p_j \mathbf{y}^p_j \ge \mathbf{y}^p_o \ \forall p \\ &\lambda^p_j \in \Omega_p, \ \forall p, \quad \beta^p \in \mathbb{R}_+ \end{aligned} \tag{9}$$

The objective function of (9) provides the optimal value of each process inefficiency (and the sum shall not be interpreted as a firm inefficiency, which we will address below). The set of constraints of this linear programme represents the production possibilities set for each production process p. This programme is in all respects a standard Directional distance function programme; and since the constraints associated with each process are all disjoint, the optimal solution of this programme is the same as solving P linear programmes separately.

The scale and convexity properties of the process technologies are given by one option in the following set (we omitted non-negativity constraints on decision variables to

save on notation):

$$\Omega_p = \left\{ \lambda_j^p \geq 0;\ \sum_{j=1}^{J} \lambda_j^p = 1;\ \sum_{j=1}^{J} \lambda_j^p \leq 1;\ \sum_{j=1}^{J} \lambda_j^p \geq 1;\ \sum_{j=1}^{J} \lambda_j^p = S;\ \lambda_j^p \in \{0,1\} \right\} \tag{10}$$

The previous options include respectively: constant returns to scale (CRS), variable returns to scale (VRS), non-increasing returns to scale (NIRS), non-decreasing returns to scale (NDRS), size efficient economies and the free disposal hull (FDH). The constraint set Ω_p is not indexed by j because we allow for it to be process specific but not firm specific. Notice that we are not assuming any particular scale or convexity assumption in what follows (though considering non-convexity means that some programmes will become MILP). Note that some authors have addressed situations where different processes may have different returns to scale characterisations, like Cook and Zhu (2011) or Hennebel et al. (2017) both in the context of multiple output technologies (related to the output-specific inputs literature).

Model (9) is very similar to model (1) (see Färe 1986 or Färe and Grosskopf 2000) and also to the multi-output models of Cherchye et al. (2013). The main difference between existing models and our model is that in model (9) a different score of inefficiency is allowed for each process (β^p), while existing models typically associate a single radial factor to all inputs of the processes and therefore the resulting score is the maximum of process inefficiencies rather than the inefficiency of each process, as shown before.

In what follows, and for decomposition purposes, we will consider the same directional vector in all assessments. Later on we will discuss on the choice of the directional vector.

Consider now the firm as a whole. The total input and output quantities of firm o are $\mathbf{X}_o$ and $\mathbf{Y}_o$ and the overall inefficiency of the firm can be determined solving model (11).

$$\begin{aligned} &\max_{\delta_o, \gamma_j^p} \delta_o \\ &\sum_{p=1}^{P} \sum_{j=1}^{J} \gamma_j^p \mathbf{x}_j^p \leq \mathbf{X}_o - \delta_o \mathbf{g} \\ &\sum_{p=1}^{P} \sum_{j=1}^{J} \gamma_j^p \mathbf{y}_j^p \geq \mathbf{Y}_o \\ &\gamma_j^p \in \Omega_p,\ \forall p, \quad \delta_o \in \mathbb{R}_+ \end{aligned} \tag{11}$$

The above firm model resembles existing ones in the literature in particular model (2) (Kao 2009a, 2012). Note however, that Kao (2009a, 2012) does not recognise the need for a prior step for measuring process efficiency nor the existence of different types of factors (allocatable and process specific). The input constraints in model (11) are equivalent to the ones of model (2) for allocatable inputs and equivalent to the ones in model (1) for process specific inputs (since in this case the aggregate input vector of process specific inputs is just the input value for one specific process as all the other values in the sum are zero, since the data matrices have been manipulated ex-ante to allow for this). Model (11) also resembles the centralised allocation models of Lozano and Villa (2004), or the more recent model of Cherchye et al. (2017), where coordination efficiency was obtained from the comparison between a centralised model similar to our (11) and a decentralised model similar to our model (9). In our case, we argue that the differences between the sum of process inefficiencies and the firm inefficiency is due to reallocation inefficiencies. Indeed, if process efficient targets (obtained from model (9)) are employed in model (11) we get:

$$\begin{aligned} &\max_{\gamma_o, \Lambda_j^p} \gamma_o \\ &\sum_{p=1}^{P} \sum_{j=1}^{J} \Lambda_j^p \mathbf{x}_j^p \leq \sum_{p=1}^{P} (\mathbf{x}_o^p - \beta_o^{*p} \mathbf{g}) - \gamma_o \mathbf{g} \\ &\sum_{p=1}^{P} \sum_{j=1}^{J} \Lambda_j^p \mathbf{y}_j^p \geq \mathbf{Y}_o \\ &\Lambda_j^p \in \Omega_p,\ \forall p, \quad \gamma_o \in \mathbb{R}_+ \end{aligned} \tag{12}$$

Where the input constraint right hand side can be re-arranged to: $\sum_p (\mathbf{x}_o^p) - \sum_p \beta_o^{*p} \mathbf{g} - \gamma_o \mathbf{g} \Longleftrightarrow \mathbf{X}_o - (\sum_p \beta_o^{*p} + \gamma_o)\mathbf{g}$. This means that overall firm inefficiency is in fact equivalent to $\delta_o = \sum_p \beta_o^{*p} + \gamma_o$.

Our first decomposition of the inefficiency of the firm is therefore shown in (13).

$$\delta_o^* = \gamma_o^* + \sum_p \beta_o^{*p}, \quad \forall o = 1, \ldots, J \tag{13}$$

Clearly, while the technical inefficiency components ($\sum_p \beta_o^{*p}$) have to do with inefficiencies arising in production at the process level, the reallocation component (γ_o^*) has to do with misallocation decisions made at the firm level and it is therefore not a type of inefficiency which can be attributed to the individual processes (since for the processes the allocation is given).

The idea of a component of system efficiency that is due to reallocation of resources can be seen in the network DEA context in Nemoto and Goto (2003), Bogetoft et al. (2009) and Färe et al. (2018). In these cases, dynamic network DEA models have been proposed and the authors propose a measure of dynamic efficiency which is related to the reallocation of intermediate factors. In parallel systems this decomposition, to the authors knowledge, has not been proposed before.

We note that model (11) identifies the firm efficiency and overall targets for the inputs and outputs of the firm: $(\mathbf{X}^*{}_o, \mathbf{Y}^*{}_o)$. Process specific targets obtained from $\sum_j \gamma_j^{*p}\mathbf{x}_j^p$ and $\sum_j \gamma_j^{*p}\mathbf{y}_j^p$ are in fact in an infinite number as shown in Asmild et al. (2009). In this paper we do not address the issue of solving the problem of multiple solutions to these models, but a procedure such as that suggested by Asmild et al. (2009) could be used here.

The next step in our analysis is to look at potential mis-allocation of resources at the industry level or, in other words, mis-allocation of resources across firms. The aggregation argument is similar to the one made from the process level to the firm level, except that here we are going to sum across processes and also across firms and consider the possibility of reallocating inputs and outputs not only across processes within a firm but also across firms themselves. Model (14) makes the industry reallocation problem explicit. Note that in this model we added an index k to the intensity variables in order to consider sums across firms. This is because we are summing all intensity variables that are associated with each firm and process in all assessments of each firm k. So contrary to the previous models, (14) is solved just once.

$$\begin{aligned} &\max_{\Upsilon_{jk}^p,\eta} \eta \\ &st \sum_{k=1}^{K}\sum_{p=1}^{P}\sum_{j=1}^{J}\Upsilon_{jk}^p\mathbf{x}_j^p \le \sum_{k=1}^{K}\sum_{p=1}^{P}\mathbf{x}_k^p - \eta\mathbf{g} \\ &\sum_{k=1}^{K}\sum_{p=1}^{P}\sum_{j=1}^{J}\Upsilon_{jk}^p\mathbf{y}_j^p \ge \sum_{k=1}^{K}\sum_{p=1}^{P}\mathbf{y}_k^p \\ &\Upsilon_{jk}^p \in \Omega_p,\ \forall p, \quad \eta \in \mathbb{R}_+ \end{aligned} \tag{14}$$

Given that the optimal solution of (9) is a feasible solution of (11) and the optimal solution of (11) is a feasible solution of (14) we have that $\eta^* > \sum_k \delta_k^* > \sum_k \sum_p \beta_k^{*p}$. As a result, the total system (or industry) inefficiency $IE = \eta^*$ can be decomposed additively into a component arising from mis-allocation of resources at the industry level $IRE = \tau^* = \eta^* - \sum_k \delta_k^*$, a component deriving from mis-allocation of resources at the firm level $FRE = \sum_k \gamma_k^*$ and a process technical inefficiency component ($PTE = \sum_k \sum_p \beta_k^{*p}$), returning the following overall decomposition:

$$IE = IRE + FRE + PTE = \tau^* + \sum_k \gamma_k^* + \sum_k \sum_p \beta_k^{*p} \tag{15}$$

The left hand side of this expression is measuring the overall input inefficiency (or excess of input use) at the system level; the right hand side is attributing this overall excess of input use to mis-allocation deriving from the system allocation (in the form of a market failure or a central planner failure), a component measuring the mis-allocation of resources within each firm and a process technical inefficiency component measuring the input excess deriving from misuse of resources during the production process. Attempts to attribute the overall inefficiency of the system (IE) to the individual production processes exhaustively would fail to grasp the difference between allocation of resources in the planning phase and use of resources during the production phase. We should also point to the fact that having an additive decomposition of input inefficiencies gives an opportunity to look at the percentage contribution of these different components onto the overall inefficiency of the system:

$$1 = \frac{IRE}{IE} + \frac{FRE}{IE} + \frac{PTE}{IE} \tag{16}$$

In fact, the percentage contribution of process p on the total inefficiency of the system is β_k^{*p}/IE and the percentage contribution of firm k reallocation will be γ_k^*/IE. This is informative on the importance of particular production processes and firms onto the overall inefficiency of the system. For example the total percentage contribution of process p for all firms can be measured as $\sum_k \beta_k^{*p}/IE$.

We note that model (14) similarly to (11) also identifies multiple process specific targets for each unit k obtained from $\sum_j \Upsilon_{jk}^{*p}\mathbf{x}_j^p$ and $\sum_j \Upsilon_{jk}^{*p}\mathbf{y}_j^p$. We direct the reader to Asmild et al. (2009) for a solution to this problem which we do not address here.

5 Our approach applied to the illustrative example

The application of the models presented above to our illustrative example results in an overall industry inefficiency of 0.3 when variable returns to scale (VRS) are employed. Since we are using a directional vector that is equal to the total input use of the industry, an inefficiency value of 0.3 means that a potential saving of 30% of the inputs would be possible. This means that the industry efficiency is 70%. The overall industry inefficiency of 0.3 can be decomposed as follows: 0.1 for industry resource reallocation (IRE), 0.118 for firm level resource reallocation (FRE) and 0.082 for processes inefficiencies (PTE). The way firm efficiency decomposes within each process can be seen in the Table 6.

Values of process inefficiencies correspond to the efficiency scores under independent assessment of processes, however in this case the values are not expressed in efficiency radial scores but inefficiency values according to the directional distance function approach. The efficiency scores of the DMUs correspond to the system technology model (2) presented before which is equivalent to our model

Table 6 Inefficiencies for each process at various levels

	Proc 1 Inefficiency	Proc 2 Inefficiency	Proc 3 Inefficiency	Total processes Inefficiency	Reallocation Inefficiency	DMU Inefficiency
DMU1	0	0.021	0.02	0.041	0.005	0.046
DMU2	0.016	0	0	0.016	0.025	0.041
DMU3	0	0	0	0	0	0
DMU4	0.025	0	0	0.025	0.088	0.113
Total	0.041	0.021	0.02	0.082	0.118	0.2

(11) for this simple case of a single allocatable input. As a result, if we take DMU4 its inefficiency of 0.113 corresponds to a reduction of observed inputs from 260 to 191.07 corresponding to an efficiency score of 73.49% (and approximately the same as in Table 3). So, in terms of efficiency measurement our approach produces consistent results to those observed in the literature (when the underlying type of inputs and outputs are the same). However, the decomposition and recognition of the existence of reallocation inefficiencies when one moves to higher hierarchical levels than the subunit level, is not common in the literature. Reallocation inefficiency implies exchanges of inputs across firms or across processes within the firm, and such movements are visible in the targets obtained from the solved models. Taking DMU4 as an example, it uses an overall input of 260 to produce an overall amount of 128 units of output 1 and 170 units of output 2. The industry model proposes some reallocations within firms. In particular it proposes that DMU4 should reduce its input consumption to 127.2 (in fact in the industry model all DMUs should reduce their input consumption except DMU1 that should increase it from 120 units to 130). The rearrangements between firms correspond to 33.3% (0.1/0.3) of the overall industry inefficiency (0.3). The remaining 66.7% correspond to within firms inefficiency for which DMU4 is the largest inefficiency contributor: (56.5% i.e., 0.113/0.2). The Firm model proposes an input target for DMU4 of 191.2 - this target implies that within DMU4 process 1 should consume 30 units, process 2 should consume 101.2 units and process 3 should consume 60 units of input. However, these values correspond largely to a reallocation of the input between processes because the observed levels were 45, 200 and 15, respectively. That is, processes 1 and 2 should consume less input, but process 3 should quadruplicate its input consumption. Clearly the process inefficiency model could not assess this reallocation, since in process inefficiency we were just looking at similar processes and see the extent to which inputs could be reduced without sacrificing outputs. For DMU4 only process 1 was considered inefficient and a target of input consumption of 30 was devised in the process inefficiency assessment. So for process 1 there are no reallocation inefficiencies identified since the process target and the firm target are the same. But for process 2 and 3 this is not so and reallocation inefficiencies are identified. In terms of input savings these reallocation contribute to further savings of (245–191.2 =) 53.8, which expressed in percentage of the total input consumption of the industry corresponds to 8.8%.

A final note to call the attention to the fact that our models can identify the degree of reallocation inefficiencies but are not built to provide optimal allocations. Clearly additional constraints could be included in the models to condition the reallocations to be obtained in the final solution, but for sake of simplicity we do not follow that avenue in this paper.

6 Empirical application to hospitals

6.1 Data on Portuguese public hospitals

We illustrate the proposed approach to data of Portuguese public hospitals in 2008. Data are provided at service level, and then aggregated at the hospital level and later on at the industry level. Only seven specialties have been considered in our analysis, but these correspond to a large proportion of the services provided by Portuguese hospitals.

Average values of the data used (for each specialty) are shown in Table 7, where we separate those variables that were considered inputs and those that were considered outputs.

In Table 7, the inputs or resources used in the specialties are human resources (number of doctors (Doc) and nurses (Nur)), beds, which can be seen as an indicator of the size of the inpatient wards, and other resources proxyed by the aggregate cost (cost) of several items (drugs, clinical material, complementary means of diagnosis, surgery ward costs and other supplies and service costs; this overall cost proxies quantity variables assuming that hospitals face similar prices Portela 2014). Outputs in Table 7 represent the main services provided by each specialty within an hospital: inpatient days of stay (Indays), outpatient appointments (OutA), hospital sessions (DaySess), and surgeries (surg), that only happen in surgical specialties—for example cardiology is not a surgical specialty. Note that in general hospital outputs (like days of stay) are adjusted

Table 7 Descriptive statistics per specialty

	Cardiology	General surgery	Internal medicine	Orthoapedics	Paediatrics	Oncology	Gynaecolhogy
N. Hospitals	29	36	36	35	29	27	33
Avg Beds	18.76	70.83	97.71	52	33.24	1.81	47.24
Avg Doc	14.24	28.2	44.14	19.11	28.24	3.70	26.58
Avg Nur	28.22	58.42	70.0	37.36	43.96	8.43	57.82
Avg Costs	5911778.46	11173612.49	8454473.37	6018385.97	3477782.8	9364737.57	7253031.01
Avg Indays	5791.10	20947.86	35488.37	14622.26	5200.56	498.41	11582.61
Avg OutA	8994.38	14670.94	10367.8	14591.51	9664	5942.78	17089.37
Avg DaySess	530.45	964.09	0	865.66	910.59	8646.04	222.7
Avg Surg	0	2630.17	0	1693.57	200.86	0	1866.33

Table 8 Descriptive statistics at the hospital level

	Average	Max	Min	St Dev
Beds	309.57	688	106	136.82
Docs	154.57	396	27	92.33
Nur	286.61	664.97	63.71	144.37
Costs	47,489,192.23	142,580,787.02	8,018,186.58	34,842,404.73
Emergcosts	27,084,598.74	51,476,316.13	4,594,908.42	12,039,148.35
Indays	91,471.086	200,624	25085	40,998.3
OutA	75,787.31	201,895	10,636	45,545.75
DaySess	9,903.51	46,816	0	9,907.84
Surg	6,249.86	16,611	892	3,662.77
Emergpatients	139,992.06	330,256	48,431	59,636.67

by case mix to reflect the severity of the patient's conditions. We did not apply this adjustment here because the comparison within specialties assures a more homogeneous case mix between patients. In addition to that, in Portugal case mix is only computed at the hospital level and not at the service level.

All outputs in Table 7 are process/specialty specific, while inputs are allocatable between services, with the exception of doctors (doctors allocated to one specialty are specialist doctors that cannot be allocated to any other specialty - but could be allocated to other hospitals if there was excess of doctors in one hospital and lack in another).

Table 7 shows that specialties vary widely in terms of the mix of resources and outputs produced. For example, general surgery and internal medicine are the services that show more beds and therefore inpatient days. Orthopaedics and gynaecologist, on the other hand, are the services with more outpatient appointments, while oncology has a very reduced average number of beds but a high number of day hospital sessions (related with chemotherapy and radiotherapy sessions that do not require staying overnight in hospital). General surgery is obviously the specialty with more surgeries performed, and non-surgical specialties (like cardiology, internal medicine and oncology do not have surgeries at all).

At the hospital level (summing all inputs and outputs across specialties), we have a total of 223 observations for specialties spread over 35 hospitals, and the statistics are shown in Table 8.

At the hospital level, apart from the inputs and outputs considered for the specialties we also considered emergency costs (emergcosts) on the input side and admissions at emergency (emergpatients) on the output side. The emergency service serves the whole hospital, and as a result emergency variables are considered joint non-allocatable inputs and outputs, which have a nature similar to public goods. For our sample of 35 hospitals (not all of which have all the 7 specialties; see Table 7) on average 139,803.16 patients were admitted in emergencies, and the cost of this service has been on average 27,084,598.74 thousands Euros. Hospitals have on average about 300 beds, 154 doctors and 286 nurses, meaning that they are not too big on average (even maximum values are not too high—note however that our sample is only aggregating a sample of services in hospitals and therefore does not reflect the real dimension of Portuguese hospitals).

Table 9 Summary inefficiencies

Process	Inefficiency 2.11	%	# 0 5
S1- cardiology	0.1697	8.04%	18
S2 - general surgery	0.432	20.48%	23
S3 - internal medicine	0.8453	40.07%	19
S4 - orthopaedics	0.2696	12.78%	19
S5 - paediatrics	0.1633	7.74%	18
S6 - oncology	0.0455	2.16%	22
S7 - gynaecology	0.1844	8.44%	20
REA_{WH}	1.46		5
Firm	3.68		5

6.2 Main results

The models discussed in the previous sections were applied to our data set, using the average industry inputs as a directional vector, and under the assumption of non-decreasing returns to scale at the process and firm level. For each hospital an inefficiency score was obtained and decomposed into services inefficiencies and reallocation inefficiencies. Gams software was used to obtain the results and the code is available upon request. Detailed results are also available upon request, since for sake of brevity here we just discuss the main results of the analysis.

In addition to efficiency scores we analysed potential savings at the various levels of decision. To do this we computed targets for inputs and outputs resulting from the solution of the three level models. For example, input targets for the process/services are obtained from the process inefficiency model (9) as $\sum_j \lambda_j^{*p} \mathbf{x}_j^p$, from the firm efficiency model (11) as $\sum_j \gamma_j^{*p} \mathbf{x}_j^p$, and from the industry model (14) as $\sum_j \Upsilon_{jk}^{*p} \mathbf{x}_j^p$.

Industry inefficiency obtained from model (14) is equal to 7.25, and this value can be decomposed into services inefficiencies (2.11), reallocation inefficiency within hospitals (1.46) and reallocation inefficiency across hospitals (3.68). As a result, potential savings accrue mainly from reallocation in the industry, followed by the elimination of service inefficiencies and finally through the elimination of reallocation inefficiencies within hospitals (REA_{WH}). Table 9 shows the sum of inefficiencies for the set of 36 hospitals, the number of efficient units in each level of analysis, and the inefficiency decomposition. Total service inefficiencies and its importance are also shown.

Process inefficiency is the sum of the service inefficiency (2.11) of all hospitals. Reallocation inefficiency within hospital is 1.46, meaning that the total hospital inefficiencies identified are 3.68. There are only 5 hospitals that are overall efficient, implying that they are both process and reallocation efficient (see in the Appendix that these hospitals are H7, H11, H17, H29 and H32). It is worth mentioning again that process efficiency only happens when all processes within the hospital are efficient.

When analysing processes inefficiencies we can see that there are many services that show small inefficiencies. For example, S6 (Oncology) has 22 units that are efficient, it contributes to the overall inefficiency of processes on average by 2.16% (see Table 9). On the contrary the Internal medicine service contributes to the overall process inefficiencies by over 40%, and this service has the highest total inefficiency value, meaning that it is the service where more potential savings can be found.

We define savings as the difference between observed and target levels, and percentage savings as the ratio between savings and industry total observed values. Savings on each input per service are shown in Fig. 2, where the sum of the three bars indicates total savings (difference between observed values and industry targets as a percentage of the observed value). All services are advised to reduce their beds in order to maximise industry efficiency - the service that has a higher potential for reducing beds is S6 (oncology). Note that this big potential implies in fact the reduction of few beds, since this service is already the one with less beds on average. Regarding doctors, in the process efficiency assessment where comparable services are benchmarked, the model identifies a large potential for reducing their number. However, this potential is offset by industry targets that rearrange services in a way that doctors are in fact required and may even be higher than observed. For example, in S1 (cardiology) the observed number of doctors and industry targets are the same and, therefore, no global savings are identified in doctors for this service. As a result, the potential savings in doctors that are identified at the process level, do not mean that doctors are to be disposed off but they are to be reallocated. The only service where reductions in doctors were still identified at the industry level was S3 (internal medicine), which was the most inefficient at process level. Regarding nurses, in most services the industry model still identifies an excess of nurses except in S5 (paediatrics). As far as the input Costs, savings were identified in most specialties, but in S5 they shall in fact increase (since the increase in costs identified at the industry level is bigger than the savings identified at process level).

The industry model implied a varied number of reallocations across hospitals that need to be looked at carefully. First it is important to notice that, given the multiple optimal solutions of the industry model, the reallocation obtained is just one in an infinite number of possibilities, that we did not address in full. Second, the NDRS constraint prevented zero targets for the inputs and outputs of the hospital services in the process and firm assessments, but it could not prevent zero targets (interpreted as the closure of that

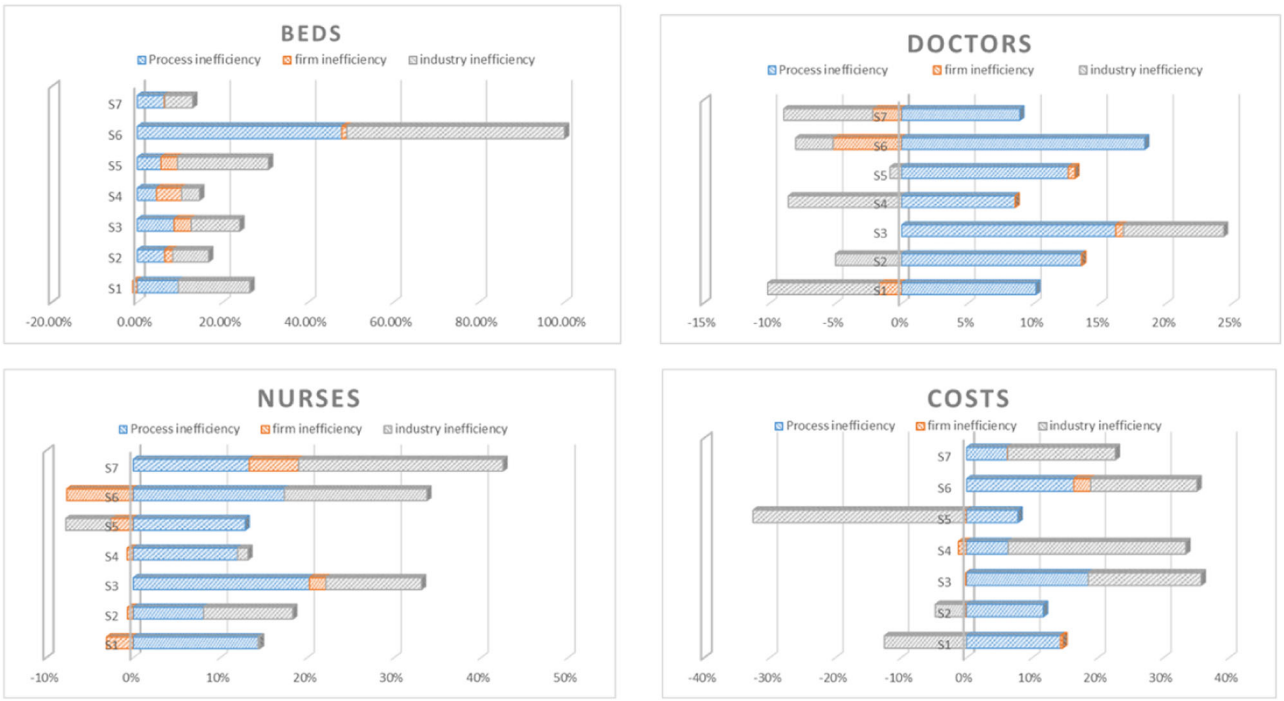


Fig. 2 Total savings per service and input, expressed as a cumulative percentage of total

Table 10 N. hospitals with each service in observed data and industry model

	Hospitals Observed	Hospitals Industry model
Cardiology	29	5
General surgery	35	35
Internal medicine	35	35
Orthoapedics	35	35
Paediatrics	29	4
Oncology	27	4
Gynaecology	33	4

service/process) for the industry assessment. Since not all hospitals have all services, zero targets may occur from zero **x** and **y** values, and not from zero intensity variables in model (14). In fact, the industry model suggests the closure of some services in some hospitals and the re-dimensioning of some services in others. Overall the number of hospitals with each of the considered services in our data is shown in Table 10. This table also shows the number of services that the industry model 'advises'.

Interestingly the industry model suggests that 4 services should be concentrated in a fewer number of hospitals. Three of these services are specialised services and indeed we have specialised hospitals for children, for oncological patients and maternities. Note that out of the 5 efficient hospitals (see Appendix) H7 does not have S6 (oncology), H11 does not have S1 (cardiology) and H17 does not have S6 and S7 (gynaecology), and as a result the industry model considers these specialties concentrated in a reduced number of hospitals. Clearly additional constraints can be imposed in the industry model to avoid extreme reallocations (e.g., geographical constraints that could impose some services to be maintained in some regions of the country). See Pachkova (2009) for some constraints on allocation and the identification of similar problems when full reallocation is allowed.

7 Conclusion

In this paper, we proposed a framework for the measurement of the inefficiency of a parallel network system and the attribution of this inefficiency to component parts. This framework is based on efficiency models computed at various levels of analysis and reconciles previous literature that has been developed mostly in an unrelated way (like network DEA models, industry models, and output specific input models). The models proposed should be seen as a first attempt to compute and decompose efficiency at the firm level into the efficiency of its processes. In this decomposition we recognise that the firm inefficiency is not just the sum of the efficiency of the component parts, since firm inefficiency involves not only technical inefficiency but also allocation inefficiency. In this paper these issues are resolved for parallel production models but we note that

extensions to other type of network models are possible and desirable. Relevant to the literature is the clarification of input and output types that should be considered when we are in the presence of network structures.This paper contributes also to this discussion and clarification.

We believe that this paper contributes to open avenues of research. Network DEA models have been represented in the literature through multiplier and envelopment models. This has been the cause of some confusion since, when the internal structure of the network is modelled, the two approaches may yield conflicting views. This has led Chen et al. (2013) to suggest that depending on the objectives of the analysis one should use one or the other form - the multiplier form when the interest in finding subunits efficiency simultaneously with DMUs efficiency, and the envelopment model when the objective is to find frontier projections (see also Castelli and Pesenti 2014). Connections with multiplier models may form the material for another paper and are not discussed here because of space constraints. All our programmes admit a dual (therefore a multiplier form) and it would be quite interesting to see the implications of our models for the dual formulation. One point that we should make clear is that the associated shadow prices one derives from the three different programmes are going to refer to shadow prices for that particular level of aggregation: for example the shadow prices associated with the process level inefficiency are going to be the process level evaluation of the value of the resources used; on the contrary the shadow prices associated with the system level are going to be the system level evaluation of the value of the resources used in each particular process. These are very different interpretations and somehow will yield irreconcilable views about the value of the resources. For example the value of one particular machine can be very low in a given process, but it can be very high for the system as a whole, pointing to the fact that this particular machine should probably be not part of the resource endowment of that specific process.

Finally, we should mention that an important area of research should be how to allow for costly reallocation. In general reallocating inputs and production across the different parts of the system may incur a cost (for example training workers to accomplish different tasks or reallocating inputs inter-temporally may be costly). Our analysis (and most of the models included in the references) can be used to give an ex-ante estimate of the potential gains from these reallocations and can be used to monitor the efficiency of the system while these reallocations are being implemented. How to include data on the cost of this reallocation is an important area of future research.

Funding Open access funding provided by FCT|FCCN (b-on).

Compliance with ethical standards

Conflict of interest The authors declare no competing interests.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

8 Appendix: Detailed results from the assessment of 36 hospitals

Table 11

Table 11 Summary inefficiencies

Hospital	Service	Process	REA_{WH}	Firm	Hospital	Service	Process	REA_{WH}	Firm
H1	S1	0.0037	0.1199	0.1324	H18	S1	0.0017	0.0646	0.1042
H1	S2	0			H18	S2	0		
H1	S3	0			H18	S3	0		
H1	S4	0			H18	S4	0.0198		
H1	S5	0			H18	S5	0.0181		
H1	S6	0			H18	S6			
H1	S7	0.0088			H18	S7	0		
H2	S1	0	0.0282	0.0603	H19	S1		0.0022	0.0072
H2	S2	0.0052			H19	S2	0		
H2	S3	0.0269			H19	S3	0.005		

Table 11 (continued)

Hospital	Service	Process	REA_{WH}	Firm	Hospital	Service	Process	REA_{WH}	Firm
H2	S4	0			H19	S4	0		
H2	S5	0			H19	S5	0		
H2	S6	0			H19	S6	0		
H2	S7	0			H19	S7	0		
H3	S1		0.0194	0.0536	H20	S1	0.0037	0.0423	0.0667
H3	S2	0			H20	S2	0		
H3	S3	0.0012			H20	S3	0.0207		
H3	S4	0.0012			H20	S4	0		
H3	S5	0.0318			H20	S5	0		
H3	S6	0			H20	S6			
H3	S7	0			H20	S7	0		
H4	S1	0	0.0129	0.0358	H21	S1		0.0494	0.0851
H4	S2	0			H21	S2	0.0315		
H4	S3	0			H21	S3	0.0042		
H4	S4	0.023			H21	S4	0		
H4	S5	0			H21	S5			
H4	S6	0			H21	S6			
H4	S7	0			H21	S7			
H5	S1	0.0066	0.0182	0.0549	H22	S1	0	0.0789	0.1385
H5	S2	0			H22	S2	0		
H5	S3	0			H22	S3	0.0504		
H5	S4	0.0207			H22	S4	0.0004		
H5	S5				H22	S5	0.0049		
H5	S6	0			H22	S6	0		
H5	S7	0.0093			H22	S7	0.0039		
H6	S1	0	0.0244	0.0252	H23	S1	0.0244	0.0146	0.1128
H6	S2	0.0008			H23	S2	0.0043		
H6	S3	0			H23	S3	0.0435		
H6	S4	0			H23	S4	0.0046		
H6	S5	0			H23	S5	0.0117		
H6	S6	0			H23	S6	0		
H6	S7	0			H23	S7	0.0098		
H7	S1	0	0	0	H24	S1	0	0.0602	0.0967
H7	S2	0			H24	S2	0.0142		
H7	S3	0			H24	S3	0.0025		
H7	S4	0			H24	S4	0		
H7	S5	0			H24	S5			
H7	S6				H24	S6			
H7	S7	0			H24	S7	0.0199		
H8	S1	0.0176	0.0038	0.0511	H25	S1	0	0.0311	0.0571
H8	S2	0			H25	S2	0		
H8	S3	0			H25	S3	0		
H8	S4	0.0297			H25	S4	0.002		
H8	S5	0			H25	S5	0.0072		
H8	S6	0			H25	S6	0		
H8	S7	0			H25	S7	0.0168		
H9	S1	0	0.0523	0.4727	H27	S1	0	0.2116	0.3551

Table 11 (continued)

Hospital	Service	Process	REA_{WH}	Firm	Hospital	Service	Process	REA_{WH}	Firm
H9	S2	0.1342			H27	S2	0		
H9	S3	0.2436			H27	S3	0.1436		
H9	S4	0.0426			H27	S4	0		
H9	S5	0			H27	S5			
H9	S6	0			H27	S6	0		
H9	S7	0			H27	S7	0		
H10	S1	0	0.0017	0.0397	H28	S1	0.0025	0.0347	0.1268
H10	S2	0.026			H28	S2	0		
H10	S3	0			H28	S3	0.0522		
H10	S4	0			H28	S4	0.0119		
H10	S5				H28	S5	0.0255		
H10	S6	0			H28	S6			
H10	S7	0.012			H28	S7	0		
H11	S1		0	0	H29	S1	0	0	0
H11	S2	0			H29	S2	0		
H11	S3	0			H29	S3	0		
H11	S4	0			H29	S4	0		
H11	S5	0			H29	S5	0		
H11	S6	0			H29	S6	0		
H11	S7	0			H29	S7	0		
H12	S1	0.0019	0.0089	0.0216	H30	S1	0	0.0553	0.1029
H12	S2	0			H30	S2	0		
H12	S3	0			H30	S3	0		
H12	S4	0			H30	S4	0.0072		
H12	S5	0.0109			H30	S5	0.0118		
H12	S6	0			H30	S6	0.0067		
H12	S7	0			H30	S7	0.0218		
H13	S1	0	0.054	0.1042	H31	S1	0.0855	0.2733	0.7439
H13	S2	0.0101			H31	S2	0.1567		
H13	S3	0			H31	S3	0.1474		
H13	S4	0.0358			H31	S4	0.0405		
H13	S5	0			H31	S5			
H13	S6	0.0025			H31	S6	0.0066		
H13	S7	0.0018			H31	S7	0.0339		
H14	S1		0.0188	0.0745	H32	S1	0	0	0
H14	S2	0.0114			H32	S2	0		
H14	S3	0.0131			H32	S3	0		
H14	S4	0.0012			H32	S4	0		
H14	S5	0.0287			H32	S5	0		
H14	S6	0			H32	S6	0		
H14	S7	0.0012			H32	S7	0		
H15	S1	0.018	0.0583	0.109	H33	S1	0	0.0008	0.0023
H15	S2	0			H33	S2	0		
H15	S3	0			H33	S3	0.0015		
H15	S4	0			H33	S4	0		
H15	S5	0.0096			H33	S5	0		
H15	S6	0.0231			H33	S6			

Table 11 (continued)

Hospital	Service	Process	REA_{WH}	Firm	Hospital	Service	Process	REA_{WH}	Firm
H15	S7	0			H33	S7	0		
H16	S1	0	0.0707	0.1143	H34	S1	0.0041	0.0292	0.1636
H16	S2	0.0078			H34	S2	0.0298		
H16	S3	0.0359			H34	S3	0.0536		
H16	S4	0			H34	S4	0.0264		
H16	S5	0			H34	S5	0.0031		
H16	S6	0			H34	S6	0.0066		
H16	S7	0			H34	S7	0.0108		
H17	S1	0	0	0	H35	S1	0	0.0042	0.0068
H17	S2	0			H35	S2	0		
H17	S3	0			H35	S3	0		
H17	S4	0			H35	S4	0.0026		
H17	S5	0			H35	S5	0		
H17	S6				H35	S6	0		
H17	S7				H35	S7	0		
					H36	S1		0.0123	0.0467
					H36	S2	0		
					H36	S3	0		
					H36	S4	0		
					H36	S5	0		
					H36	S6	0		
					H36	S7	0.0344		

References

Afsharian M, Ahn H, Harms S. G(2021) A review of DEA approaches applying a common set of weights: The perspective of centralized management. Euro J Operat Res 294(1):3–15

Asmild M, Paradi JC, Pastor JT (2009) Centralized resource allocation bcc models. Omega 37:40–49

Banker R (1992) Selection of efficiency evaluation models. Contemp Account Res 9:343–355

Beasley J (1995) Determining teaching and research efficiencies. J Oper Res Soc 46:441–452

Bogetoft P, Färe R, Grosskopf S, Hayes K, Taylor L (2009) Dynamic network dea: an illustration (< special issue> operations research for performance evaluation). J Oper Res Soc Jpn 52:147–162

Castelli L, Pesenti R (2014) Network, shared flow and multi-level DEA models: a critical review. In: Cook W, Zhu J (eds) Data Envelopment Analysis, International Series in Operations Research and Management Science, Springer, New York, 208, 329–376

Castelli L, Pesenti R, Ukovich W (2010) A classification of DEA models when the internal structure of the decision making units is considered. Ann Oper Res 173:207–235

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

Chen Y, Cook W, Kao C, Zhu J (2013) Network dea pitfalls: Divisional efficiency and frontier projection under general network structures. Eur J Oper Res 226:507–515

Cherchye L, Rock BD, Dierynck B, Roodhooft F, Sabbe J (2013) Opening the black box of efficiency measurement: Input allocation in multi-output settings. Oper Res 61:1148–1165

Cherchye L, Rock BD, Hennebel V (2017) Coordination efficiency in multi-output settings: a dea approach. Ann Oper Res 250:205–233

Cherchye L, Rock BD, Walheer B (2015) Multi-output efficiency with good and bad outputs. Eur J Oper Res 240:872–881

Cook W, Chai D, Doyle J, Green R (1998) Hierarchies and groups in DEA. J Prod Anal 10:177–198

Cook W, Green R (2004) Multicomponent efficiency measurement and core business identification in multiplant firms: a DEA model. Eur J Oper Res 157:540–551

Cook W, Green R (2005) Evaluating power plant efficiency: a hierarchical model. Comput Oper Res 32:813–823

Cook WD, Hababou M, Tuenter HJ (2000) Multicomponent efficiency measurement and shared inputs in data envelopment analysis: an application to sales and service performance in bank branches. J Prod Anal 14:209–224

Cook WD, Zhu J (2011) Multiple variable proportionality in data envelopment analysis. Oper Res 59:1024–1032

Despić O, Despić M, Paradi JC (2007) DEA-R: Ratio-based comparative efficiency model, its mathematical relation to DEA and its use in applications. J Prod Anal 28:33–44

Färe R (1986) A dynamic non-parametric measure of output efficiency. Oper Res Lett 5:83–85

Färe R, Grabowski R, Grosskopf S, Kraft S (1997) Efficiency of a fixed but allocatable input: a non-parametric approach. Econ Lett 56:187–193

Färe R, Grosskopf S (1996) Intertemporal production frontiers: with dynamic DEA, Kluwer Academic Publishers, Boston

Färe R, Grosskopf S (1996) Productivity and intermediate products: a frontier approach. Econ Lett 50:65–70

Färe R, Grosskopf S (2000) Network dea. Soc Econ Plan Sci 34:35–49

Färe R, Grosskopf S, Li S-K (1992) Linear programming models for firm and industry performance. Scand J Econ 94:599–608
Färe R, Grosskopf S, Margaritis D (2010) Time substitution with application to data envelopment analysis. Eur J Oper Res 206:686–690
Färe R, Grosskopf S, Margaritis D, Weber WL (2018) Dynamic efficiency and productivity. In: Grifell-Tatjé E, Lovell CK, Sickles RC (eds) The Oxford handbook of productivity analysis, Oxford University Press, Oxford, p 183–210
Färe R, Grosskopf S, Whittaker G (2007) Network dea. In: Zhu J, Cook W (eds) Modelling data irregularities and structural complexities in data envelopment analysis, Springer, p 209–240
Farrell MJ (1957) The measurement of productive efficiency. J R Stat Soc A 120:253–281
Forsund FR, Hjalmarsson L (1979) Generalised Farrell measures of efficiency: an application to milk processing in Swedish dairy plants. Econ J 89:294–315
Golany B, Phillips F, Rousseau J (1993) Models for improved effectiveness based on dea efficiency results. IIE Trans 25:2–10
Golany B, Tamir E (1995) Evaluating efficiency-effectiveness-equality trade-offs: a data envelopment analysis approach. Manag Sci 41:1172–1184
Hennebel V, Simper R, Verschelde M (2017) Is there a prison size dilemma? an empirical analysis of output-specific economies of scale. Eur J Oper Res 262:306–321
Johansen L (1972) Production functions; an integration of micro and macro, short run and long run aspects, North-Holland Publishing Company
Kantorovich LV (1960) Mathematical methods in the organization and planning of production. Leningr Univ Engl Transl Manag Sci 6:4
Kao C (2009) Efficiency decomposition in network data envelopment analysis: a relational model. Eur J Oper Res 192:949–962
Kao C (2009) Efficiency measurement for parallel production systems. Eur J Oper Res 196:1107–1112
Kao C (2012) Efficiency decomposition for parallel production systems. J Oper Res Soc 63:64–71
Kao C (2013) Dynamic data envelopment analysis: a relational analysis. Eur J Oper Res 227:325–330
Kao C (2014) Network DEA analysis: a review. Eur J Oper Res 239:1–16
Kao C (2016) Efficiency decomposition and aggregation in network data envelopment analysis. Eur J Oper Res 255:778–786
Kao C (2017) Network Data Envelopment Analysis, 2nd edn. Springer
Kao C (2018) Multiplicative aggregation of division efficiencies in network data envelopment analysis. Eur J Oper Res 270:328–336
Kao C, Hwang S-N (2010) Efficiency measurement for network systems: IT impact on firm performance. Decis Support Syst 48:437–446
Karagiannis G (2015) On structural and average technical efficiency. J Prod Anal 43:259–267
Koopmans TC (1951) An analysis of production as an efficient combination of activities. In: Koopmans TC (ed) Activity Analysis of Production and Allocation, Proceeding of a Conference, John Wiley and Sons Inc, London, p. 33–97
Kuosmanen T, Cherchye L, Sipilainen T (2006) The law of one price in data envelopment analysis: restricting weight flexibility across firms. Eur J Oper Res 170:735–757
Li S, Cheng Y (2007) Solving the puzzles of structural efficiency. Eur J Oper Res 180:713–722
Lim S, Zhu J(2016) A note on two-stage network DEA model: Frontier projection and duality. Eur J Oper Res 248(1):342–346
Lozano S (2011) Scale and cost efficiency analysis of networks of processes. Expert Syst Appl 38:6612–6617
Lozano S, Villa G (2004) Centralized resource allocation using data envelopment analysis. J Prod Anal 22:143–161
Lozano S, Villa G, Adenso-Dıaz B (2004) Centralised target setting for regional recycling operations using dea. Omega 32:101–110
Mar Molinero C (1996) On the joint determination of efficiencies in a data envelopment analysis context. J Oper Res Soc 47:1273–1279
Nemoto J, Goto M (1999) Dynamic data envelopment analysis: modeling intertemporal behavior of a firm in the presence of productive inefficiencies. Econ Lett 64:51–56
Nemoto J, Goto M (2003) Measurement of dynamic efficiency in production: an application of Data Envelopment Analysis to Japanese electric utilities. J Prod Anal 19:191–210
Pachkova EV (2009) Restricted reallocation of resources. Eur J Oper Res 196:1049–1057
Peyrache A (2013) Industry structural inefficiency and potential gains from mergers and break-ups: a comprehensive approach. Eur J Oper Res 230:422–430
Peyrache A (2015) Cost constrained industry inefficiency. Eur J Oper Res 247:996–1002
Peyrache A, Silva MC (2022) A comment on decomposition of efficiency in network production models. CEPA Working Paper Series
Peyrache A, Silva MC (2022) Efficiency and productivity analysis from a system perspective: Historical overview. In: Chotikapanich D, Rambaldi AN, Rohde N (eds) Advances in Economic Measurement: A Volume in Honour of DS Prasada Rao, Springer Nature Singapore, Singapore
Podinovski V, Olsen O, Sarrico C (2018) Nonparametric production technologies with multiple component processes. Oper Res 66:282–300
Podinovski VV (2022) Variable and constant returns-to-scale production technologies with component processes. Oper Res 70(2):1238–1258
Portela M (2014) Value and quantity data in economic and technical efficiency measurement. Econ Lett 124:108–112
Portela M et al. (2016) Benchmarking hospitals through a web based platform. Benchmarking Int J 23:722–739
Salerian J, Chan C (2005) Restricting multiple-output multiple-input dea models by disaggregating the output–input vector. J Prod Anal 24:5–29
Shephard RW, Färe R (1980) Dynamic theory of production correspondences, vol. 50, Verlag Anton Hain
Silva MCA (2018) Output-specific inputs in dea: an application to courts of justice in portugal. Omega 79:43–53
Ylvinger S (2000) Industry performance and structural efficiency measures: Solutions to problems in firm models. Eur J Oper Res 121:164–174