Data Envelopment Analysis Journal, 2018, 4: 109–147

The Remarkable Incidence of Congestion in Production: A Review, Empirical Illustration, and Research Agenda

Kristiaan Kerstens[1] and Ignace Van de Woestyne[2*]

[1] *IÉSEG School of Management, CNRS-LEM (UMR 9221), Univ. Lille, 3 rue de la Digue, F-59000 Lille, France; k.kerstens@ieseg.fr*
[2] *KU Leuven, Research unit MEES, Warmoesberg 26, B-1000 Brussel, Belgium*

ABSTRACT

This contribution surveys the existing economic literature measuring congestion using nonparametric specifications of technologies. The focus is on the magnitude and especially the incidence of the congestion detected using traditional radial input-oriented efficiency measures. Furthermore, it shows the limitations of this radial measurement and how alternative measurement schemes may reveal higher amounts of congestion. This is empirically illustrated using a variety of secondary data sets (which guarantees replicability of our results).

Keywords: Nonparametric technology; congestion; DEA

*We thank the editor and two constructive referees for their helpful comments that have led to substantial improvements. The usual disclaimer applies.
Corresponding Author: Kristiaan Kerstens.

ISSN 2161-1823; DOI 10.1561/103.00000026

1 Introduction

Few studies have empirically documented the phenomenon of congestion in production, here intuitively defined as production where marginal productivity has become negative. One of the largest streams in the literature where at least some studies document congestion employs multi-output nonparametric production technologies that impose either ray or free disposability to distinguish between technical inefficiency (understood as production below the production frontier) and congestion (for the moment defined as a particular severe form of technical inefficiency). While the empirical analysis of efficiency and productivity has become quite popular (see, e.g., Badunenko and Romero-Ávila, 2013; Henderson and Russell, 2005), congestion is most often ignored in such studies, despite the fact that some studies find it to be the most important source of poor performance (e.g., Zhengfei and Oude Lansink, 2003).

Traffic congestion is a prominent example. Throughout the world, all major cities suffer from severe congestion as manifested by reduced speeds and traffic flows over a given road network. Another well-documented example is output loss in agriculture due to excessive use of fertilisers. Agronomic crop response models relating crop yield to nutrients in general show a maximal plateau (where marginal product of input is zero) but also a phase where crop yield declines (where marginal product of input has become negative). The latter is sometimes denoted as the toxic range (e.g., Jones, 2001, pp. 216–221).

There is a limited axiomatic literature allowing to reveal and measure some limited forms of congestion in production.[1] In this contribution, we provide an empirical perspective on these limited forms of congestion in production. In particular, this paper has two goals. First, we want to document the amounts and incidence of congestion that are empirically observed in the available literature and to systematically illustrate these amounts and incidence using several secondary data sets under some limited variations in assumptions on technology. Second, we want to see how the way one measures congestion affects the amounts and incidence

[1] These forms of congestion are known as monotone output-limitational (MOL) congestion (see Färe and Svensson, 1980).

that are revealed and also this is systematically illustrated using these same secondary data sets.

This paper is structured as follows. Section 2 provides some basic definitions of technology and its boundaries. Furthermore, it discusses the representation of technologies by means of efficiency measures and introduces the nonparametric technologies used in the empirical part of this paper. Section 3 introduces the distinction between technical inefficiency and congestion measurement as part of some well-known static efficiency decompositions. In a second subsection, we review the empirical literature containing some evidence on the amounts and incidence of congestion. In a final subsection, we illustrate how the traditional radial way of measuring efficiency and congestion actually may underestimate the amounts of congestion. Furthermore, we outline an alternative approach that does not share this defect with the radial measure. Thereafter, we present an empirical Section 4 revisiting several existing data sets and exploring the amounts and incidence of congestion while contrasting the radial efficiency measure as well as the alternative approach. Section 5 develops a systematic research agenda. Section 6 concludes.

2 Technologies: Definition, Subsets, and Representation

2.1 Technology: Definition, Subsets, and Representation

A production technology describes all possibilities how input vectors $x = (x_1, \ldots, x_m) \in \mathbb{R}^m_+$ can be transformed into output vectors $y = (y_1, \ldots, y_n) \in \mathbb{R}^n_+$. The technology T summarises the set of all feasible input and output vectors: $T = \{(x, y) \in \mathbb{R}^{m+n}_+ : x \text{ can produce } y\}$. Given the focus on efficiency measurement in the input orientation, technology can be represented by the input correspondence $L : \mathbb{R}^n_+ \to 2^{\mathbb{R}^m_+}$, where $L(y)$ is the set of all input vectors that yield at least the output vector y:

$$L(y) = \{x : x \text{ can produce } y\}. \tag{1}$$

It is useful to distinguish between three subsets of the input set $L(y)$ denoting production units on the boundary. First, one can define the isoquant of an input set as:

$$Isoq\, L(y) = \{x \in L(y) : \lambda x \notin L(y), \forall \lambda \in [0, 1[\}. \tag{2}$$

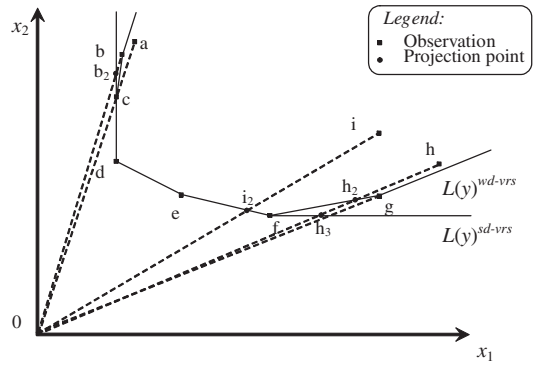

Figure 1: Input set and its subsets.

Second, the weak efficient subset is defined by:

$$WEff\ L(y) = \{x \in L(y) : u < x \Rightarrow u \notin L(y)\}. \tag{3}$$

Finally, the efficient subset of an input set is defined as:

$$Eff\ L(y) = \{x \in L(y) : u \leq x \text{ and } u \neq x \Rightarrow u \notin L(y)\}. \tag{4}$$

Obviously, $Eff\ L(y) \subseteq WEff\ L(y) \subseteq Isoq\ L(y) \subseteq L(y)$. For nonparametric technologies, these subsets are described in Section 2.2 and illustrated in Figure 1.

Technologies can be characterised using distance functions, which are related to the efficiency measures defined in Farrell (1957). The input-oriented Farrell efficiency measure $E^i(x, y)$ indicates the minimum contraction of an input vector by a scalar λ while still remaining part of the input set:

$$E_i(x, y) = \inf_{\lambda} \{\lambda : \lambda x \in L(y), \lambda \geq 0\}. \tag{5}$$

Obviously, $E_i(x, y) \leq 1$ for $x \in L(y)$, with unity indicating efficiency.

2.2 Nonparametric Technologies: Definitions and Subsets

Consider a set of K observations $A = \{(x_1, y_1), \ldots, (x_K, y_K)\} \in \mathbb{R}_+^{n+m}$ on the base of which we reconstruct a technology. Nonparametric

specifications of technology can be estimated by enveloping this set of observations while maintaining some basic production axioms.

First, under variable returns to scale (VRS) we define both a weakly and a strongly disposable technology. Under strong input and output disposal (SD), a variable returns to scale technology is defined as:

$$L(y|SD,VRS)=\left\{x:x\geq\sum_{k=1}^{K}z_kx_k,y\leq\sum_{k=1}^{K}z_ky_k,\sum_{k=1}^{K}z_k=1,z\geq 0\right\}. \tag{6}$$

The vector z represents the activity variables that indicate the intensity at which a particular activity is employed in constructing the reference technology. Under weak input disposal (WD), strong output disposal, and variable returns to scale the technology is defined as:

$$L(y|WD,VRS)=\left\{x:\gamma x=\sum_{k=1}^{K}z_kx_k,y\leq\sum_{k=1}^{K}z_ky_k,\sum_{k=1}^{K}z_k=1,\gamma\in(0,1],z\geq 0\right\}. \tag{7}$$

Note that the inequalities on the input dimensions have now been replaced by an equality while the observed inputs are scaled down by the scalar γ.

Second, under constant returns to scale (CRS) we can equally define both a weakly and a strongly disposable technology by simply removing the constraint $\sum_{k=1}^{K}z_k=1$ in the technologies (6) and (7), respectively. These technologies are denoted $L(y|SD,CRS)$ and $L(y|SD,CRS)$, respectively. Details on these technologies and the underlying axioms are found in Hackman (2008) or Ray (2004).

Figure 1 shows typical isoquants for such nonparametric input sets with strong disposability of inputs and with weak disposability of inputs starting from some basic observations. Note that the weakly disposable technology (7) is normally a subset of the strongly disposable technology (6). We now clarify the three subsets ((2)–(4)) on these input sets. For both technologies, the efficient subset *Eff* $L(y)$ consists of the line segments joining points *def*. For the weakly disposable technology, the weakly efficient subset *WEff* $L(y)$ contains the connected line segments

cdef, and its isoquant *Isoq* $L(y)$ is formed by adding the line segments bc and fg to those in *WEff* $L(y)$. Points on the rays through $0b$ and $0g$ belong to the boundary of the input correspondence, not to any of its three subsets. For the technology with strong disposability, the weakly efficient subset and the isoquant coincide: both contain the connected line segments *cdef* and the lines beyond c and f parallel to both axes.

3 Technical Inefficiency and Congestion: Framework and Empirical Perspective on Congestion Measures and Incidence

3.1 Static Efficiency Decomposition: The Role of Congestion

The distinction between technical efficiency (TE) and structural efficiency (STE) or congestion can be seen against the background of a variety of proposals to develop a static taxonomy of efficiency. The seminal article by Farrell (1957) distinguished between technical and allocative inefficiency. Seitz (1970) was the first to add a further distinction by defining a scale efficiency component based on cost function comparisons. Later on Førsund and Hjalmarsson (1974), Färe *et al.* (1983b), and Banker *et al.* (1984) distinguished between technology-based technical and scale efficiency, whereby the second team of authors also integrated a congestion component. Färe *et al.* (1985c) were among the first to offer an extended efficiency decomposition summarising most of the above developments.[2]

Since the focus is on congestion measurement, we first intuitively explain the notion of congestion and how it can be revealed with the help of the technologies defined above using Figure 1. To illustrate how the weakly disposable technology models congestion, we start from observation f. While the strongly disposable technology allows to waste additional inputs x_1 at no opportunity cost, the weakly disposable technology leaves two options: either the wasting of extra inputs x_1 requires additional costs in terms of extra inputs x_2 to reach, for instance, observation g while maintaining current output levels, or the wasting of extra inputs x_1 without any additional inputs x_2 results in reaching another input set of the weakly disposable technology with a lower level

[2]These decompositions (including the distinction between short- and long-run measures as well as the integration of capacity measures) are surveyed in De Borger *et al.* (2012).

of outputs. In short, wasting additional inputs x_1 has an opportunity cost in terms of either additional inputs x_2 or less outputs.

Then, we define and illustrate the traditional radial way of measuring technical efficiency and congestion as proposed in Färe *et al.* (1983b). The radial measure of input congestion can be defined as follows:

$$C_i(x, y|\, VRS) = \frac{E_i(x, y|SD,\, VRS)}{E_i(x, y|WD,\, VRS)}, \tag{8}$$

where $E_i(x, y|SD, VRS)$ and $E_i(x, y|WD, VRS)$ denote the radial efficiency measure (5) defined relative to VRS technologies with strong disposability (6) and weak disposability (7), respectively. Since $E_i(x, y|SD, VRS) \leq E_i(x, y|WD, VRS)$, the ratio $C_i(x, y|\, VRS) \leq 1$.

This leads to the following decomposition of pure technical efficiency:

$$E_i(x, y|SD, VRS) = E_i(x, y|WD, VRS) \cdot C_i(x, y|\, VRS). \tag{9}$$

The left-hand side is the pure technical efficiency measure $E_i(x, y|SD, VRS)$ evaluated with respect to a technology with strong disposability. On the right-hand side we have a weak technical efficiency measure $E_i(x, y|WD, VRS)$ evaluated with respect to a technology with weak disposability times the congestion measure $C_i(x, y|\, VRS)$ as defined in (8). This whole decomposition is measured with respect to VRS technologies.

This radial congestion measure $C_i(x, y|\, VRS)$ can be illustrated by commenting on observation h situated in the interior of the input set $L(y|WD, VRS)$ in Figure 1. Weak technical efficiency $E_i(x, y|WD, VRS)$ is represented by the ratio of distances $0h_2/0h$ measured relative to the input set $L(y|WD, VRS)$. Pure technical efficiency $E_i(x, y|SD, VRS)$ is represented by the ratio of distances $0h_3/0h$ relative to h_3 on the weakly efficient subset ($WEff\, L(y|SD, VRS)$). Structural efficiency or congestion $C_i(x, y|\, VRS)$ is measured by the ratio of distances $0h_3/0h_2$ derived by comparing radial distances between an activity without congestion at point h_3 on the weakly efficient subset ($WEff\, L(y|SD, VRS)$) and activity with congestion h_2 on the boundary of $L(y|WD, VRS)$. Hence, using (9), the total deviation from the strongly disposable technology captured by pure technical efficiency can be decomposed into a weak technical efficiency and a congestion component.

Turning now to a comparison of observations a and b on the same Figure 1, we obtain the following results. Since observation a is projected onto the weakly efficient subset of both the strongly and weakly disposable technologies, it does not suffer from congestion but the ratio of distances $0c/0a$ is just interpreted as technical inefficiency solely. By contrast, applying the same logic, observation b is situated on the $Isoq\, L(y)$ of the weakly disposable technology and hence technically efficient. However, the gap between the strongly and weakly disposable technologies (i.e., $0b_2/0b$) reveals congestion. Noticing that observation a wastes more of both inputs than observation b for identical outputs, one may wonder why the latter is considered congested but technically efficient, while the former is technically inefficient but uncongested. We return to this issue below.

Crucial for our focus on congestion measurement in the remainder are the following remarks. First, we consider congestion as an extreme and unacceptable form of technical efficiency. While technical inefficiency is costly and implies a waste of resources, one can imagine certain reasons justifying its existence (e.g., slack resources and capacity in anticipation of an increasing demand over a product life cycle). However, congestion implies a waste of resources and an additional opportunity cost in terms of additional inputs or wasted outputs. Therefore, it is almost impossible to justify and ideally requires prompt managerial action.[3]

Second, one should clearly distinguish between detecting congestion and summarising its relative importance as a source of inefficiency within some efficiency decomposition. While the radial efficiency measure (5) is convenient to summarise the relative importance of different efficiency components in a multiplicative decomposition, as illustrated above it need not necessarily be an accurate tool to reveal the incidence of congestion (see also infra).

Third, in view of the previous argument, congestion is traditionally measured with respect to a nonparametric technology with flexible (i.e., variable) returns to scale. Färe *et al.* (1985c) distinguish between private

[3]Ray (2004, p. 184) states in this respect: "A general note of caution is strongly warranted at this point. Presence of input congestion is quite unlikely in behavioral data. Even though the marginal productivity of an input could eventually become negative, it is difficult to imagine a producer actually using the input at that level — especially when it has to be procured at a cost."

and social goals when discussing the rationale behind their decomposition components. These authors consider scale issues a social goal, while they deem technical efficiency and congestion private matters. However, some authors have implicitly or explicitly proposed alternative measurement schemes. For instance, McDonald (1996) proposes measuring congestion by contrasting weakly and strongly disposable technologies under constant returns to scale. His numerical examples illustrate how this change in returns to scale assumption affects the amount of congestion. In our view, given our research question, it is important to understand how the amounts of congestion may vary depending on the axioms maintained on technologies.[4] However, we consider the variable returns to scale technology to be the true technology, while the constant returns to scale technology is just an auxiliary technology useful to determine, e.g., the returns to scale for individual production units.[5] Therefore, this case of measuring congestion relative to constant returns to scale technologies is just included for the sake of completeness.

Therefore, we also define a radial measure of input congestion with respect to CRS technologies as follows:

$$C_i(x, y|CRS) = \frac{E_i(x, y|SD, CRS)}{E_i(x, y|WD, CRS)}, \tag{10}$$

[4]We dissent from the argument made by Färe and Grosskopf (2000) in reply to McDonald (1996) who simply refer to economic tradition to justify measuring congestion relative to a variable returns to scale technology.

[5]The issue seems to be that some proportionality between inputs and outputs seems mistakenly taken as evidence of a constant returns to scale technology. Scarf (1994, pp. 114–115) aptly ridicules the possibility of a constant returns to scale technology as follows: "Both linear programming and the Walrasian model of equilibrium make the fundamental assumption that the production possibility set displays constant or decreasing returns to scale; that there are no economies associated with production at a high scale. I find this an absurd assumption, contradicted by the most casual of observations. Taken literally, the assumption of constant returns to scale in production implies that if technical knowledge were universally available we could all trade only in factors of production, and assemble in our own backyards all of the manufactured goods whose services we would like to consume. If I want an automobile at a specified future date, I would purchase steel, glass, rubber, electrical wiring and tools, hire labor of a variety of skills on a part — time basis, and simply make the automobile myself. I would grow my own food, cut and sew my own clothing, build my own computer chips and assemble and disassemble my own international communication system whenever I need to make a telephone call, without any loss of efficiency. Notwithstanding the analysis offered by Adam Smith more than two centuries ago, I would manufacture pins as I needed them."

where $E_i(x, y|SD, CRS)$ and $E_i(x, y|WD, CRS)$ denote the radial efficiency measure (5) defined relative to technologies $L(y|SD, CRS)$ and $L(y|WD, CRS)$, respectively. Since $E_i(x, y|SD, CRS) \leq E_i(x, y|WD, CRS)$, the ratio $C_i(x, y|CRS) \leq 1$.

Now, this results in the following alternative decomposition of pure technical efficiency:

$$E_i(x, y|SD, CRS) = E_i(x, y|WD, CRS).C_i(x, y|CRS). \quad (11)$$

The left-hand side is the pure technical efficiency measure $E_i(x, y|SD, CRS)$ evaluated with respect to a technology with strong disposability. On the right-hand side we have a weak technical efficiency measure $E_i(x, y|WD, CRS)$ evaluated with respect to a technology with weak disposability times the congestion measure $C_i(x, y|CRS)$ as defined in (10). This whole decomposition is measured with respect to CRS technologies.

These same static decompositions of efficiency have also been integrated into the decompositions of productivity indices and indicators. Examples of decompositions of the efficiency change component of the Malmquist productivity index are found in Fukuyama and Weber (1999), Glass *et al.* (1997), Glass and McKillop (2000), and McCallion *et al.* (2000), or Ng and Li (2009), among others.

Note that congestion occurs only when either $C_i(x, y|VRS) < 1$ or $C_i(x, y|CRS) < 1$. Therefore, congestion incidence is somehow affected by rounding rules determining whether a congestion measure is situated below unity. We round numbers up at three decimals in the empirical part: this should normally lead to a very conservative estimate of congestion and its incidence.

3.2 Congestion Measurement: Amounts and Incidence Reported in the Empirical Literature

While congestion is widely cited as a theoretical possibility in most microeconomics textbooks, empirical evidence as to its prevalence is relatively rare. The merit of the literature applying this nonparametric efficiency decomposition outlined above is that quite a lot of studies have reported on (parts of) these efficiency decompositions, though relatively few report on congestion.

Congestion measured using either decomposition (9) or (11) is the most important source of inefficiency at the sample level in at least eight articles we are aware of: Byrnes and Färe (1987) and Byrnes *et al.* (1988) both analyse US surface coal mines, Çakmak and Zaim (1992), Wu *et al.* (2003) and Zhengfei and Oude Lansink (2003) assess Turkish, American and Dutch agriculture respectively, Färe *et al.* (1989) analyse US electric utilities, Mulumba *et al.* (2017) assess Ugandan referral hospitals, and Odeck (2006) evaluates the Norwegian public bus companies.[6] Just to offer some basic idea of the amount of waste involved, Table 1 summarises for each study the average amount of congestion efficiency as well as its incidence (% of sample). The last

Table 1: Congestion efficiency and incidence: Literature review.

Article	Congestion efficiency	Congestion incidence	Remarks
Byrnes and Färe (1987)	0.71	26.3%	$N = 186$
Byrnes *et al.* (1988)	0.74	69.0%	$N = 84$, Interior states
	0.70	83.3%	$N = 54$, Interior states; UMWA[†]
	0.77	74.3%	$N = 113$, Western states
	0.43	83.3%	$N = 12$, Western states; Nonunion
Çakmak and Zaim (1992)	0.92	38.1%	Sample
Färe *et al.* (1989)	0.925	NA[‡]	$N = 23$, Year 1969
	0.924	NA	$N = 23$, Year 1975
Mulumba *et al.* (2017)	0.921	53.8%	$N = 13$, Year 2012
	0.952	53.8%	$N = 13$, Year 2013
Odeck (2006)	0.89	57.6%	$N = 33$
Wu *et al.* (2003)	0.92	44.9%	$N = 147$
Zhengfei and Oude Lansink (2003)	0.88	75.0%[§]	$N = 1072$

[†]UMWA = affiliation with United Mine Workers of America.
[‡]NA = Not available.
[§]Text states: "approximately 3/4 of observations" (p. 475).

[6]The study of Habibullah *et al.* (2005) reports an output-oriented congestion measure: while the definition correctly defines the congestion measure as being larger than or equal to unity, the reported empirical congestion measure is smaller than or equal to unity. However, the scale efficiency measure is both theoretically and empirically correctly defined. If the empirical congestion measure just needs inverting, then congestion is again the most important source of poor performance.

column adds the sample size and some remarks whenever needed. Note that the second, fourth, and fifth studies have several entries: in the second article a basic distinction is made between the Interior and Western US states; the fourth study compares two distinct years; and the fifth article reports two out of three years for which congestion is most important. Furthermore, for the second study we also report results for those subsamples for which congestion efficiency is the key component. Remark that we have been scanning for studies reporting the highest congestion inefficiencies rather than the studies reporting the highest congestion incidence levels relative to the incidence of other sources of poor performance.

Several conclusions can be drawn from Table 1. First, congestion inefficiency can vary from a modest 7.5% (=1 − 0.925) to a high 29% (=1 − 0.71) at the sample level. In the second study, for Western nonunion mines one even observes a staggering 57% (=1 − 0.43) congestion inefficiency. Second, the incidence of congestion inefficiency varies widely: between a low 26.3% to about 75% of the sample. For the second study, two subsamples even record an incidence of 83.3%. Finally, congestion inefficiency and incidence need not be correlated. For instance, the lowest incidence coincides with the highest congestion inefficiency (see Byrnes and Färe, 1987). By contrast, the second lowest congestion inefficiency goes hand in hand with the highest incidence levels (Byrnes *et al.*, 1988). In particular, modest congestion inefficiency levels can hide high incidence levels (see Odeck, 2006; Zhengfei and Oude Lansink, 2003). In brief, these studies reveal a wide variety of patterns of congestion inefficiency and incidence, even though the sample sizes of most studies are quite modest.

Adding further evidence on the low inefficiency high incidence combination, one striking example is found in a comparative study of French and US hospitals reported in Dervaux *et al.* (2004). Congestion affects 79.8% and 54.4% of French and US hospitals for rather modest congestion levels that amount to 0.968 and 0.966, respectively. Sample size of French and US hospitals is 1,080 and 903, respectively. Other studies with smaller sample size often yield quite similar results. For instance, Färe *et al.* (1985a) indicate that 57.5% of electric utilities are congested though congestion efficiency is 0.947 on average, Kritikos *et al.* (2010) list that 70.9% of Greek telecommunication branches suffer from congestion which amounts to 0.954 on average, while Puig-Junoy (2000)

reports a congestion incidence of 40.4% for a level of 0.972. Sample sizes of the last three studies are 151, 127, and 94, respectively. Finally, Habibullah *et al.* (2005) study 37 commercial banks listed on the Kuala Lumpur Stock Exchange in the period 1988–1993: in four out of six years congestion incidence is highest compared to the incidence of other sources of inefficiency, while in the two remaining years it is at a tie with the incidence of other sources of inefficiency: it varies between a low 56.8% and a high 83.8%.

Furthermore, while in some studies congestion inefficiency does not dominate at the sample level, it may well turn out to be important for specific parts of the sample. We again offer a selection of illustrations from the literature known to us. Byrnes *et al.* (1987) document that congestion dominates for Illinois grain farms smaller than 700 acres, representing 72.9% of the sample. Färe *et al.* (1987) study US electric utilities from the western states at the plant level over three years. When aggregating results over plants and years they find that for three out of nine firms congestion is the main source of inefficiency. When aggregating results over years, 7 out of 22 plants have a main problem with congestion. Flegg *et al.* (2004) analyse 45 British universities over the academic years 1980–1981 till 1992–1993: for 8 out of 13 periods congestion is the main source of underperformance.[7] In a similar vein, Flegg and Allen (2009b) study 41 new British universities (i.e., former polytechnics becoming universities in 1992) over the period 1995–1996 to 2003–2004: congestion is now the main source of bad performance for eight out of nine periods. Glass and McKillop (2000) report that the change in input congestion is the most important source of efficiency change in the input Malmquist index on average in the subperiod 1991–1993.

Finally, there is the possibility that congestion plays a negligible role at the sample level or for specific parts of it, but that it is critically important for some particular observations. For instance, evaluating British building societies in 1985 and finding scale inefficiency as the prime source of underperformance, Field (1990) observes that congestion is most important for about 9.9% of observations (with amounts

[7]The same article also traces performance over time using a Malmquist productivity index and its components, including a congestion component: these static results are duplicated in the dynamic results.

between 0.48 and 0.78). Simões and Marques (2011) also report that scale inefficiency dominates for 68 Portuguese hospitals. Nevertheless, congestion inefficiency is critical for three hospitals (with amounts between 0.57 and 0.86). Zeitsch and Lawrence (1996) analyse 10 base load power plants over 7 time periods: focusing on a single period, congestion is the dominant source of technical inefficiency for 4 plants (with amounts between 0.86 and 0.96). Field and Emrouznejad (2003) find congestion to be crucial for 4 out of 22 Scottish neonatal care units in the period 1993–1994. Finally, re-analysing data on the Chinese automobile and textile industries in the period 1981–1997, Flegg and Allen (2009a) obtain in an intertemporal analysis that congestion is critical for both sectors in a single year over this time horizon (1989 and 1995 for automobile and textile industries, respectively). Mulumba *et al.* (2017) report that for the year 2012 congestion dominates for 5 out of 13 observations. Nasierowski and Arcelus (2003) find that the innovation system in about 30% of countries suffers from congestion and that it is a dominant source for some of these countries. Finally, Glass *et al.* (1997) observe that the change in output congestion is the crucial source of efficiency change in the Malmquist index for one and three university departments in the subperiods 1989–1992 and 1992–1996, respectively.

Notice that in all of the above except one, we have limited ourselves to studies defining congestion in terms of a gap between strongly and weakly disposable technologies with variable returns to scale (i.e., decomposition (9)). When imposing stronger assumptions (e.g., when congestion is measured relative to strongly and weakly disposable technologies under constant (instead of variable) returns to scale), congestion results may well worsen. For instance, Simões and Marques (2011) report that average congestion inefficiency increases from 0.962 to 0.955 when imposing variable and constant returns to scale, respectively. In a similar vein, Flegg and Allen (2009b) indicate that congestion remains the main source of under-performance for eight out of nine periods, but the congestion efficiency component under constant returns to scale is lower in every single year compared to the one evaluated relative to variable returns to scale.

Thus, from this literature review it is difficult to deny that congestion may well play a serious role as a source of poor performance in a relatively wide range of sectors. Furthermore, the sometimes high incidence of

congestion seems to indicate that in these samples a lot of observations are situated close to the isoquant and boundary of the input set, and not that close to the weak or strong efficient subset as it is implicitly or explicitly assumed. At first sight, this seems to imply that the amounts of congestion measured are not artifacts created by just a few outlying observations enveloped by a particular axiomatic structure imposed on technology.

3.3 Congestion Measurement: Limitations of Radial Measures and Remedies[8]

These amounts of congestion incidence are surprising if one realises that the use of radial efficiency measures may actually underestimate the prevalence of congestion. This is easy to illustrate with the use of Figure 2. Only observations outside the cone spanned by *WEff* $L(y|WD, VRS)$ can be subject to congestion when using radial efficiency measures: this cone is represented by the rays Oc and Of in Figure 2. Let us compare, for example, points g and i in Figure 2. Point g is revealed as being congested, since it is efficient relative to $L(y|WD, VRS)$ but not relative to $L(y|SD, VRS)$. However, point i, which is identical to g in its use of x_1 but using a higher amount of x_2 is not subjected to congestion, since the radial efficiency measure projects observation i onto the efficient subset at point i_2. Thus, the

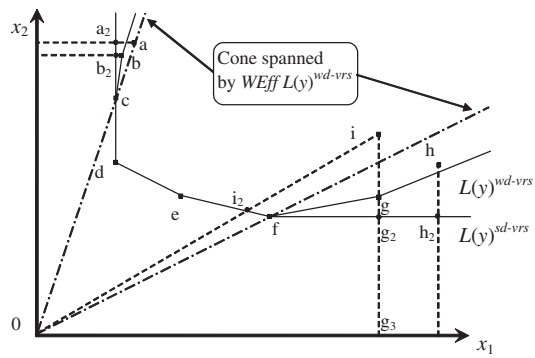

Figure 2: Limitations of radial congestion measurement.

[8]This subsection builds and extends upon Briec *et al.* (2018, pp. 2944–2945).

traditional radial way of measuring congestion may well underestimate its empirical amounts and/or its incidence.[9]

Assuming that one is willing to accept the argument that one should distinguish between the detection of congestion and summarising its relative importance as a source of inefficiency within some kind of static efficiency decomposition, then it is easy to understand that some authors have proposed to measure congestion in a nonradial way. Indeed, Färe *et al.* (1983b, p. 187) already stressed that their proposal to measure congestion radially in essence is a way to capture a nonradial phenomenon: input congestion emerges from the excessive usage of one or a subset of inputs, and it need not affect all inputs simultaneously.

One candidate solution is to measure congestion per specific input dimension. This procedure is illustrated on Figure 2 for observation i, that remained undetected using the traditional radial efficiency measure. By contrast, measuring in the direction of the second input allows detecting its congesting excessive utilisation of inputs. In particular, the distances g_3g/g_3i and the ratios of distances g_3g_2/g_3g $[= (g_3g_2/g_3i)/(g_3g/g_3i)]$ measure the amount of technical efficiency and congestion in the direction of the second input respectively. Similarly, observation b which remained uncongested using the radial measure may now be detected as being congested in the direction of the first input. Note that in addition to this component-wise approach, Dervaux *et al.* (1998) propose several nonradial and almost nonradial decompositions to summarise the relative importance of the congestion component.

To make this procedure introduced by Dervaux *et al.* (1998) explicit, we need to introduce an asymmetric efficiency measure that just looks for reductions in a single input. This asymmetric efficiency measure goes back to Färe (1975) and can be defined as follows:

$$AF_i^j(x,y) = \inf_{\lambda_j} \{\lambda_j : (x_1,\ldots,\lambda_j x_j,\ldots,x_m) \in L(y), \lambda_j \geq 0\} \quad \text{with } j \in \{1,\ldots,m\}. \tag{12}$$

Several remarks are in order. First, $AF_i^j(x,y) \leq 1$, with unity indicating efficiency. Second, there is an obvious relation with the radial efficiency measure (5): $AF_i^j(x,y) \leq E_i(x,y)$. Third, while the relation between

[9] As pointed out by a referee, this analysis may require qualification in higher dimensions.

the radial efficiency measure (5) and the asymmetric efficiency measure (12) can be signed, the input-oriented congestion measures (8) based on a ratio of such two different measures cannot be signed. Fourth, measuring efficiency using this definition (12) relative to the technologies $L(y|SD, VRS)$ and $L(y|WD, VRS)$ leads to a nonlinear programming problem. However, as shown in the Appendix, these programming problems can be linearised.[10] Note that the survey of input-oriented efficiency measures of Russell and Schworm (2009) ignores this asymmetric efficiency measure.

Several studies are known to us that have implemented such unidimensional measurement scheme for congestion: Färe *et al.* (1985b), Flegg and Allen (2009b), Fukuyama (1997), and Zhengfei and Oude Lansink (2003). For the sake of brevity, we look into the details of just one of these studies. Focusing on the Zhengfei and Oude Lansink (2003) study, while the radial input efficiency measure evaluated over all eight input dimensions leads to on average an amount of 11.7% congestion inefficiency, the use of a subvector measure per input dimension separately leads to average congestion inefficiency levels at the sample level from a minimum of 22.1% for "Other variable inputs" to a maximum of 45.6% for "Nitrogen fertiliser". Using the radial input efficiency measure, the incidence of congestion is about 75%. The use of subvector measures per input dimension leads to incidence levels varying between a minimum 35% for "Other pesticides" to 59% for "Nitrogen fertiliser". One conclusion is obviously that the radial way of measuring congestion may underestimate the amounts of congestion inefficiency relative to an input-specific measurement scheme. Its effect on congestion incidence is clearly not clear-cut.

Finally, more refined measurement schemes have been developed looking for subsets of dimensions responsible for congestion (see, for instance, Byrnes *et al.*, 1988; Färe *et al.*, 1994; Flegg and Allen, 2009b). We deliberately ignore further issues in the recent literature like, for instance, the use of directional distance functions and how the choice of direction vectors may well affect congestion measurement (see Davutyan *et al.*, 2014). Therefore, it remains somewhat an open issue how to

[10] Dervaux *et al.* (1998) do not provide these programming problems. The programming problems mentioned in the empirical application of Zhengfei and Oude Lansink (2003, p. 471) are incorrect. Therefore, for the sake of clarity, we develop these programming problems explicitly.

best measure congestion efficiency and its incidence: radially, unidimensionally, or still in some other way.

4 Empirical Illustration

In this section, we first briefly introduce the data sets adopted from existing studies. Then, we present empirical results on technical efficiency and especially congestion. We report descriptive statistics and the level of incidence.

4.1 Secondary Data Sets Employed

To empirically illustrate these developments, we employ several existing data sets that are publicly available.[11] Table 2 summarises some key features of each data set: sample size, number of inputs and outputs, and the sector. These data sets have been sorted in Table 2 according to their sample size. In the other tables we maintain this same order.

The main points to note are the following. There are three single output samples, and one multiple-output sample. One sample is from agriculture, two from industry, and one from a service sector. Sample sizes vary from rather small to rather big. There is one small unbalanced panel (Färe *et al.*, 1983a) and three cross-sections (Atkinson and

Table 2: Empirical data sources.

Article	Sample	#Inputs	#Outputs	Sector	Remarks
Färe *et al.* (1983b)	86	3	1	Electricity	Unbalanced ($T = 5$)
Porembski *et al.* (2005)	142	2	11	Banking	
Atkinson and Dorfman (2009)	192	3	1	Electricity	Year of monthly data
Fan *et al.* (1996)	471	3	1	Agriculture	

[11] Färe *et al.* (1983a) publish their data in Table 2 (pp. 358–359). The data in Porembski *et al.* (2005) are available upon request from these authors. The data from Atkinson and Dorfman (2009) and Fan *et al.* (1996) are available from the *Journal of Applied Econometrics* and *Journal of Business & Economic Statistics* archives, respectively.

Dorfman, 2009; Fan *et al.*, 1996; Porembski *et al.*, 2005).[12] Note that the time dimension in the panel data set is ignored: this amounts to assuming that there is no technical change over the five time periods.

4.2 Descriptive Statistics

Basic descriptive statistics at the sample level for the decompositions of pure technical efficiency under VRS (9) and under CRS (11) are reported in Table 3. The following conclusions can be drawn with regard to the amounts and incidence of pure and weak technical efficiency and congestion using traditional radial efficiency measures. First, as to their relative amounts, the amount of technical inefficiency is always larger on average than congestion for a given returns to scale assumption. Second, one can observe that technical inefficiency increases when moving from VRS to CRS, while there is no clear relation when comparing congestion under VRS and CRS. Third, the incidence of technical inefficiency is always higher than that of congestion. Fourth, while the incidence of technical inefficiency increases when moving from VRS to CRS, there is no such relation when comparing the incidence of congestion under VRS and CRS. Finally, the variation in incidence across samples is larger for congestion than for technical inefficiency.

Note that these results may well be affected by our rounding rules. For instance, for the Färe *et al.* (1983a,b) article we obtain an incidence of 0.919 under CRS when rounding at three decimals. Thus, from the 86 observations, 79 units are congested and only 7 are uncongested. However, when we round at six decimals, then only four observations are uncongested. This issue of rounding when measuring congestion probably deserves further investigation (but, it may equally so affect the measurement of scale efficiency and the like).

Furthermore, we have also tested for the differences in distribution between technical efficiency and congestion relative to VRS versus CRS technologies. We employ a formal test statistic proposed by Li (1996),

[12]In fact, Atkinson and Dorfman (2009) is an unbalanced panel of monthly data between April 1986 to December 1997. However, we just employ the single year 1997 for which the panel is balanced. By ignoring the time dimension over this single year, we assume that technical change can be safely ignored.

Table 3: Technical inefficiency and congestion: Amounts and incidence using a radial efficiency measure (5).

		VRS			CRS		
Sample		$E_i(x,y\|SD,.)$	$E_i(x,y\|WD,.)$	$C_i(x,y\|.)$	$E_i(x,y\|SD,.)$	$E_i(x,y\|WD,.)$	$C_i(x,y\|.)$
Färe *et al.* (1983b)	% Ineffic. Obs.	0.791	0.698	0.721	0.953	0.860	0.919
	Geom. Mean	0.928	0.946	0.981	0.894	0.919	0.973
	St. Dev.	0.056	0.052	0.029	0.063	0.057	0.045
	Min.	0.808	0.813	0.863	0.708	0.757	0.708
	Li-test†				3.235	3.421	5.416
Porembski *et al.* (2005)	% Ineffic. Obs.	0.585	0.570	0.113	0.746	0.746	0.014
	Geom. Mean	0.883	0.886	0.996	0.821	0.821	1.000
	St. Dev.	0.130	0.130	0.013	0.143	0.144	0.000
	Min.	0.502	0.502	0.919	0.445	0.445	0.995
	Li-test†				4.675	5.168	0.593
Atkinson and Dorfman (2009)	% Ineffic. Obs.	0.833	0.734	0.464	0.922	0.854	0.385
	Geom. Mean	0.741	0.777	0.954	0.632	0.658	0.961
	St. Dev.	0.203	0.207	0.092	0.229	0.236	0.092
	Min.	0.132	0.132	0.374	0.065	0.088	0.072
	Li-test†				4.480	8.013	0.936
Fan *et al.* (1996)	% Ineffic. Obs.	0.896	0.805	0.401	0.962	0.917	0.501
	Geom. Mean	0.802	0.820	0.979	0.757	0.776	0.975
	St. Dev.	0.115	0.122	0.051	0.111	0.120	0.051
	Min.	0.386	0.386	0.603	0.382	0.382	0.652
	Li-test†				6.528	11.499	3.405

†Li-test critical values: 2.33 at 1% level (***); 1.64 at 5% level (**); 1.28 at 10% level (*).

which is valid for both dependent and independent variables.[13] The null hypothesis of this Li-test states that both distributions are equal for a given efficiency score and its underlying returns to scale assumption. The test statistics are reported in the last line of each part of Table 3. While pure and weak technical efficiency clearly differ under VRS and CRS, congestion levels only differ significantly under VRS and CRS for two out of four databases.

While the incidence of technical inefficiency is in line with widespread results in the frontier efficiency literature and the recognition of the huge heterogeneity in firm-level performance measures elsewhere (see, e.g., Syverson, 2011), the incidence of congestion is a bit baffling and seems to have gone unnoticed in the literature. It varies between a negligible 1.4% to a staggering 91.9%. Thus, these numbers are perfectly in line with our review of the evidence in Subsection 3.2.

In addition, we also report the results of the component-wise approach using the asymmetric efficiency measure (12) applied to each input dimension separately. Tables 4 to 7 contain the empirical results for the four data sets. The structure of these tables is identical to the previous one. We do not discuss these tables separately, but try to draw some general conclusions.

The following conclusions can be drawn with regard to the amounts and incidence of technical efficiency and congestion using traditional radial efficiency measures in contrast to the asymmetric efficiency measure (12). First, while the incidence of technical inefficiency is equal or higher compared to the radial measurement, the incidence of congestion is on average equal or substantially higher compared to the radial measurement, though sometimes it is also lower for a particular input dimension. For instance, looking at the Porembski *et al.* (2005) data set, one notices that the incidence of technical inefficiency remains constant under VRS and CRS, but the incidence of congestion moves from 11.4% under VRS to between 16.9% and 32.4%, and under CRS from just 1.4% to between 18.3% and 26.8%. Second, compared to the radial measurement technical efficiency remains sometimes constant but it also decreases quite often, though sometimes it also increases for

[13]Note that efficiency measures based on frontier estimators are not independent: efficiency levels depend, among others, on sample size, the number of input and output dimensions, etc.

Table 4: Technical inefficiency and congestion per input in Färe *et al.* (1983b): Amounts and incidence using the asymmetric efficiency measure (12).

		VRS			CRS		
Input		$E_i(x,y\|SD,.)$	$E_i(x,y\|WD,.)$	$C_i(x,y\|.)$	$E_i(x,y\|SD,.)$	$E_i(x,y\|WD,.)$	$C_i(x,y\|.)$
Input 1	% Ineffic. Obs.	0.814	0.756	0.791	0.953	0.907	0.953
	Geom. Mean	0.554	0.678	0.817	0.276	0.438	0.631
	St. Dev.	0.056	0.056	0.154	0.063	0.063	0.194
	Min.	0.808	0.808	0.260	0.708	0.708	0.101
	Li-test†				12.763	6.982	10.668
Input 2	% Ineffic. Obs.	0.814	0.721	0.744	0.953	0.860	0.942
	Geom. Mean	0.926	0.951	0.974	0.892	0.927	0.963
	St. Dev.	0.194	0.194	0.035	0.185	0.185	0.050
	Min.	0.430	0.430	0.821	0.230	0.230	0.708
	Li-test†				2.822	2.719	5.177
Input 3	% Ineffic. Obs.	0.791	0.744	0.721	0.953	0.895	0.930
	Geom. Mean	0.673	0.730	0.922	0.600	0.659	0.910
	St. Dev.	0.055	0.055	0.085	0.063	0.063	0.119
	Min.	0.808	0.808	0.604	0.708	0.708	0.227
	Li-test†				3.024	2.415	3.579

†Li-test critical values: 2.33 at 1% level (***); 1.64 at 5% level (**); 1.28 at 10% level (*).

Table 5: Technical inefficiency and congestion per input in Porembski *et al.* (2005): Amounts and incidence using the asymmetric efficiency measure (12).

		VRS			CRS		
Input		$E_i(x,y\|SD,.)$	$E_i(x,y\|WD,.)$	$C_i(x,y\|.)$	$E_i(x,y\|SD,.)$	$E_i(x,y\|WD,.)$	$C_i(x,y\|.)$
Input 1	% Ineffic. Obs.	0.585	0.570	0.324	0.746	0.746	0.183
	Geom. Mean	0.837	0.854	0.980	0.741	0.745	0.994
	St. Dev.	0.169	0.158	0.044	0.188	0.185	0.020
	Min.	0.410	0.475	0.715	0.409	0.420	0.828
	Li-test†				4.702	5.687	1.543
Input 2	% Ineffic. Obs.	0.585	0.585	0.169	0.746	0.746	0.268
	Geom. Mean	0.764	0.767	0.995	0.680	0.685	0.993
	St. Dev.	0.218	0.215	0.015	0.221	0.219	0.019
	Min.	0.307	0.322	0.883	0.266	0.309	0.863
	Li-test†				3.899	3.881	0.593

†Li-test critical values: 2.33 at 1% level (***); 1.64 at 5% level (**); 1.28 at 10% level (*).

Table 6: Technical inefficiency and congestion per input in Atkinson and Dorfman (2009): Amounts and incidence using the asymmetric efficiency measure (12).

		VRS			CRS		
Input		$E_i(x,y\|SD,.)$	$E_i(x,y\|WD,.)$	$C_i(x,y\|.)$	$E_i(x,y\|SD,.)$	$E_i(x,y\|WD,.)$	$C_i(x,y\|.)$
Input 1	% Ineffic. Obs.	0.849	0.771	0.589	0.922	0.854	0.365
	Geom. Mean	0.457	0.550	0.831	0.323	0.351	0.921
	St. Dev.	0.322	0.319	0.226	0.316	0.336	0.158
	Min.	0.045	0.046	0.187	0.045	0.045	0.072
	Li-test†				5.752	9.604	6.759
Input 2	% Ineffic. Obs.	0.839	0.766	0.641	0.953	0.901	0.839
	Geom. Mean	0.536	0.584	0.919	0.334	0.441	0.758
	St. Dev.	0.298	0.302	0.129	0.293	0.312	0.245
	Min.	0.056	0.056	0.315	0.001	0.048	0.001
	Li-test†				14.882	10.531	4.795
Input 3	% Ineffic. Obs.	0.880	0.833	0.458	0.953	0.891	0.323
	Geom. Mean	0.433	0.515	0.841	0.309	0.395	0.782
	St. Dev.	0.310	0.315	0.212	0.283	0.310	0.243
	Min.	0.019	0.019	0.251	0.001	0.019	0.004
	Li-test†				5.162	4.715	1.508

†Li-test critical values: 2.33 at 1% level (***); 1.64 at 5% level (**); 1.28 at 10% level (*).

Table 7: Technical inefficiency and congestion per input in Fan *et al.* (1996): Amounts and incidence using the asymmetric efficiency measure (12).

		VRS			CRS		
Input		$E_i(x,y\|SD,.)$	$E_i(x,y\|WD,.)$	$C_i(x,y\|.)$	$E_i(x,y\|SD,.)$	$E_i(x,y\|WD,.)$	$C_i(x,y\|.)$
Input 1	% Ineffic. Obs.	0.896	0.817	0.709	0.962	0.928	0.900
	Geom. Mean	0.608	0.661	0.920	0.544	0.601	0.905
	St. Dev.	0.136	0.136	0.096	0.127	0.127	0.093
	Min.	0.368	0.368	0.461	0.364	0.364	0.349
	Li-test†				5.778	9.563	39.673
Input 2	% Ineffic. Obs.	0.896	0.815	0.807	0.962	0.928	0.938
	Geom. Mean	0.484	0.603	0.803	0.392	0.524	0.748
	St. Dev.	0.132	0.132	0.181	0.128	0.128	0.192
	Min.	0.379	0.379	0.160	0.366	0.366	0.140
	Li-test†				7.581	10.517	15.846
Input 3	% Ineffic. Obs.	0.896	0.851	0.688	0.962	0.936	0.771
	Geom. Mean	0.630	0.661	0.953	0.572	0.609	0.939
	St. Dev.	0.133	0.133	0.070	0.122	0.122	0.077
	Min.	0.330	0.330	0.598	0.321	0.321	0.516
	Li-test†				5.516	6.299	4.729

†Li-test critical values: 2.33 at 1% level (***); 1.64 at 5% level (**); 1.28 at 10% level (*).

a particular input dimension. However, structural efficiency tends to decrease, sometimes rather substantially. For example, looking again at the Porembski *et al.* (2005) data set, technical efficiency declines from 0.834 to between 0.799 and 0.857 under VRS, and from 0.834 to between 0.746 and 0.765 under CRS, while structural efficiency declines from 0.997 to between 0.982 and 0.995 under VRS, and from 1.000 to between 0.993 and 0.995 under CRS.

Furthermore, we have again tested for the differences in distribution between technical efficiency and congestion relative to VRS versus CRS technologies. Again, pure and weak technical efficiency clearly differ under VRS and CRS. Now, congestion levels differ significantly under VRS and CRS for two out of four databases, and only marginally so for the other two databases.

Interpreting congestion as a particular extreme form of technical inefficiency, it is hard to come up with reasons for its widespread prevalence. In the particular case of fertiliser and pesticides in agriculture, some research suggests that the excessive use relative to agronomic optimal amounts is due to uncertainty, perception biases, and the fact that the cost of over-application is low compared to the cost of under-application (e.g., see Rajsic and Weersink, 2008). In general, we are unaware of research having revealed any common causes for the observed amounts of congestion and their incidence.

5 Research Agenda

We have found in the published literature evidence that congestion is a major source of inefficiency and that the incidence of congestion can be very substantial. This hypothesis was confirmed in our own empirical analysis. It is now time to think about developing a research agenda in order to corroborate or reject our current findings in future research. We develop this agenda around a few topics. The way in which these methodological refinements may affect the amount of congestion and its incidence remains to be explored in detail: some may detract from the issue; some may amplify the phenomenon.

A first issue is that congestion is a special form of technical inefficiency that can seem hard to justify at all. However, it is important to realise that also technical inefficiency in itself — while abundantly

estimated and its amounts and incidence being accepted in part of the literature — is hard to justify at all. We develop two reflections on this.

A first reflection is about how one can think the existence of technical inefficiency when firms are supposedly operating under a high degree of competition. While traditionally technical inefficiency is conceived as incompatible with competitive markets, the framework developed by Allais (1977) and later on reformulated by Luenberger (1995) at least allows to think of the dynamics of market exchanges out of equilibrium and it considers Walrasian equilibria as limiting states where inefficiencies in consumption and production converge to zero. In this view, inefficiencies and surpluses in the economy determine the dynamics of exchange and the battle to extract surpluses. If workers and firms have some monopoly power, then workers can bargain low effort and high wages while firms strive for high effort and low wages (see, e.g., Haskel and Sanchis (2000) for such a bargaining model). In addition, incomplete contracts, poor incentives, and a variety of other explanations have been invoked to explain the existence of technical inefficiency (see, e.g., the survey in Frantz (1988), among others).

A second reflection centers on how one can think about the causes of congestion in general and for specific industries in particular. The existence of congestion is sometimes related to the law of diminishing returns: this has been presented as both a law and a statistical regularity. In agriculture, crop response models relating crop yields to essential single nutrients or combinations thereof reveal an initial phase with positive marginal product, then the existence of a maximum plateau with zero marginal product, and finally a declining phase with negative marginal products (in soil science, the latter phase is called the toxic range of nutrients).[14] In hospitals, simulation models have determined a variety of causes contributing to facility congestion (e.g., poor scheduling practices (e.g., Johnson and Happ, 1977)), congested emergency departments due to bottlenecks in long-term care facility (Patrick, 2011, etc.). From this scant evidence from the agricultural and hospital sectors, it is clear that while the existence of the congestion phenomenon is beyond doubt, its causes seem to be industry specific.

[14]For soil science: see Jones (2001, pp. 216–221), while for agricultural economics: see the survey in Paris (2008).

A second issue is that we have so far limited ourselves to some limited form of congestion. This form of congestion is known as monotone output-limitational (MOL) congestion (see Färe and Svensson, 1980). But, following Färe *et al.* (1987), the presence of congestion can also be interpreted as a violation of the weak disposability assumptions. Observations that are inefficient with respect to a weakly disposable technology then simply suggest a lack of fit between the data and the weak disposal assumption. To the extent that the goodness-of-fit with the weak disposal assumption is low this may lead to the search for alternative and weaker axioms and resulting technology specifications yielding an even closer fit with the data. When there is an upper bound to the wasting of inputs in certain directions, then one can model "hypercongestion" phenomena leading to the complete destruction of outputs (known as output prohibitive (OP) congestion (Färe and Svensson, 1980)). Briec *et al.* (2016) develop a new axiomatic approach allowing for the definition of more general multi-output technologies capable of revealing all congestion concepts, including "hypercongestion". In fact, these authors introduce a weaker axiom of S-disposability — a kind of limited strong disposal — that allows to model more general forms of congestion. A first empirical application of this new approach is found in Briec *et al.* (2018). The impact of these new models on the amount of congestion and its incidence remains to be explored.

A third issue is that by focusing on the input space solely, we have ignored the issue of measuring congestion and its incidence in the output space or in the input–output space. There are three related reflections to this.

A first reflection is that we could have used a more general weakly disposable technology. While our technology (7) assumes weak input disposal and strong output disposal, there is also a technology with both weak input and output disposal (e.g., Färe *et al.*, 1985c, p. 128) though it has rarely been empirically applied.

A second reflection is that we have not established a link with a related literature in operational research focusing on alternative approaches to measure congestion without necessarily invoking the axiom of weak disposability (see, e.g., Cooper *et al.* (1996) for the seminal alternative proposal and Kao (2010) for a recent overview). As

summarised by Kao (2010), these alternative approaches are distinct in terms of their focus on input space, output space, or input–output space. The relation to the above literature on measuring "hypercongestion" remains to be explored.

A third and final reflection is that we have limited ourselves to measuring congestion on a technology with weak input disposal solely using radial input-oriented efficiency measures on the one hand and a nonradial, asymmetric input-oriented efficiency measure on the other hand. Intuitively, it is clear that there is potentially a link between the specification of the technology allowing for congestion and the choice of orientation of measurement for some corresponding efficiency measure. A priori one could think that one should ideally evaluate congestion and its incidence in input–output space with the help of a so-called non-oriented efficiency measures in the full space of inputs and outputs.[15] In the recent literature, for instance, there is some discussion on the use of directional distance functions and how the choice of direction vectors affect congestion measurement (e.g., Davutyan *et al.* (2014) or the discussion in Briec *et al.* (2016, 2018)).

A fourth issue is that our technology (7) assuming weak input disposal and strong output disposal may well not be specified in an axiomatically correct way. In a series of articles it has been argued that this traditional specification does not satsify the convexity axiom (see, e.g., Mehdiloozad and Podinovski (2018) for a recent statement).

A fifth issue is that our technology (6) assuming strong input and output disposal may well contribute to miss-classifying non-congested observations as congested if the boundary of the input set is estimated with no or few interior facets. The reader is referred to Thrall (1996) for the definitions of interior and exterior facets and to Olesen and Petersen (2015) for an up-to-date analysis.[16]

A sixth and final issue is that our specifications of technologies may suffer from small sample bias and that no proper statistical test seems to exist to evaluate congestion. Olesen and Petersen (2016) summarise the recent literature on statistical inference for nonparametric frontier models and Kneip *et al.* (2016) provide an example of how to test

[15]For a survey of these non-oriented efficiency measures: see Russell and Schworm (2011).

[16]We thank both reviewers for making this basic point.

specific production hypotheses like convexity or constant returns to scale. It is urgently needed to have such a test for congestion and its incidence.

6 Conclusions

This contribution has offered an empirical perspective on the amounts and incidence of congestion as found in the applied frontier estimation literature. This paper has started out by describing the axiomatic production literature leading to the definition of a radial input-oriented congestion measure. This measure is capable to capture the differences between a strongly and a weakly disposable technology, whereby the latter allows for "backward bending" isoquants. Starting from studies for which congestion was the main source of inefficiency, some conclusions on the incidence of congestion at the sample level were drawn. Additional studies illustrating the importance of congestion incidence for parts of a sample and for specific individual observations were also discussed. Having discussed the intrinsic limitations of a radial measurement scheme for detecting congestion, it was proposed to detect in addition the presence of congestion and to measure its incidence using an asymmetric efficiency measure.

In the empirical section, four secondary data sets have been analysed. We apply both the traditional radial input-oriented congestion measure and the one based on the asymmetric efficiency measure. The key result is that very substantial to almost staggering congestion levels and incidence is found in these data. Furthermore, imposing constant rather than variable returns to scale does have a significant impact on the distributions of congestion levels in some, but not all, cases. Therefore, both in the empirical literature surveyed and in the samples analysed, one cannot but conclude from the congestion incidence levels that a lot of observations are situated closer to the isoquant and boundary of the input set than to the weak or strong efficient subsets.

We end with an additional concluding remarks for future research. First, it is well known that if monotonicity conditions are violated, then second-order conditions for optimizing behavior are not satisfied and duality relations break down (see, for instance, Barnett, 2002).[17] For

[17]Note that violations of curvature conditions have the same impact.

instance, Sauer (2006) surveys eight parametric frontier studies from agriculture and finds that only three articles fulfill monotonicity in all inputs. Violations of monotonicity occur as follows: for a single input in two studies; for two inputs in two articles; and even for five out of eight inputs in one study. One obvious candidate to explain these monotonicity failures is the existence of congestion. Therefore, it seems like a good idea to employ nonparametric tests of congestion in conjunction with tests of monotonicity conditions within a parametric framework to explore this potential relationship in more depth.

Second, when congestion is confirmed to be an issue, then a possible managerial implication for this finding is that this inefficiency should be removed as soon as possible, since currently it cannot be justified by any theoretical argument. Obviously, the proviso applies that our theoretical understanding of congestion as well as technical inefficiency for that matter is still in its infancy.

Appendix: Computing the Asymmetric Efficiency Measure

The computation of the asymmetric efficiency measure (12) for the evaluated observation (x_o, y_o) relative to the technology $L(y|WD, VRS)$ in (7) leads to the following nonlinear programming problem:

$$
\begin{aligned}
AF_i^j(x_o, y_o) = \min_{\theta,\delta,z_k}\ & \theta \\
\text{subject to}\ & \sum_{k=1}^{K} y_{kn} z_k \geq y_{on} && n = 1, \ldots, N, \\
& \sum_{k=1}^{K} x_{km} z_k = \theta\delta x_{om} && m = j, \\
& \sum_{k=1}^{K} x_{km} z_k = \delta x_{om} && m \in \{1, \ldots, M\}\backslash\{j\}, \\
& \sum_{k=1}^{K} z_k = 1, \\
& z_k \geq 0, \theta \geq 0, \delta \in (0, 1] && k = 1, \ldots, K.
\end{aligned}
\tag{A1}
$$

Note that this model is nonlinear: it may be difficult to linearise. Instead, we can write the problem as follows:

$$
\begin{aligned}
AF_i^j(x_o, y_o) = & \min_{\theta,\gamma,z_k} \theta \\
\text{subject to } & \sum_{k=1}^{K} y_{kn} z_k \geq y_{on} && n = 1, \ldots, N, \\
& \sum_{k=1}^{K} \gamma x_{km} z_k = \theta x_{om} && m = j, \\
& \sum_{k=1}^{K} \gamma x_{km} z_k = x_{om} && m \in \{1, \ldots, M\} \backslash \{j\}, \\
& \sum_{k=1}^{K} z_k = 1, \\
& z_k \geq 0, \theta \geq 0, \gamma \geq 1 && k = 1, \ldots, K.
\end{aligned}
\tag{A2}
$$

This model is nonlinear, but it can be linearised: put $z'_k = \gamma z_k$, then $z_k = z'_k/\gamma$. This leads to the following linear program:

$$
\begin{aligned}
AF_i^j(x_o, y_o) = & \min_{\theta,\gamma,z'_k} \theta \\
\text{subject to } & \sum_{k=1}^{K} y_{kn} z'_k \geq \gamma y_{on} && n = 1, \ldots, N, \\
& \sum_{k=1}^{K} x_{km} z'_k = \theta x_{om} && m = j, \\
& \sum_{k=1}^{K} x_{km} z'_k = x_{om} && m \in \{1, \ldots, M\} \backslash \{j\}, \\
& \sum_{k=1}^{K} z'_k = \gamma, \\
& z'_k \geq 0, \theta \geq 0, \gamma \geq 1 && k = 1, \ldots, K.
\end{aligned}
\tag{A3}
$$

Computing the same asymmetric efficiency measure (12) relative to the constant returns to scale technology $L(y|WD, CRS)$ amounts to dropping the following constraint from the above linear programming problem (A3):

$$\sum_{k=1}^{K} z'_k = \gamma.$$

References

Allais, M. (1977), "Theories of General Economic Equilibrium and Maximum Efficiency", in *Equilibrium and Disequilibrium in Economic Theory*, ed. G. Schwödiauer, Dordrecht: Reidel, 129–201.

Atkinson, S. and J. Dorfman (2009), "Feasible Estimation of Firm-Specific Allocative Inefficiency through Bayesian Numerical Methods", *Journal of Applied Econometrics*, 24(4), 675–97.

Badunenko, O. and D. Romero-Ávila (2013), "Financial Development and the Sources of Growth and Convergence", *International Economic Review*, 54(2), 629–63.

Banker, R., A. Charnes, and W. Cooper (1984), "Some Models for Estimating Technical and Scale Inefficiencies in Data Envelopment Analysis", *Management Science*, 30(9), 1078–92.

Barnett, W. (2002), "Tastes and Technology: Curvature is Not Sufficient for Regularity", *Journal of Econometrics*, 108(1), 199–202.

Briec, W., K. Kerstens, and I. Van de Woestyne (2016), "Congestion in Production Correspondences", *Journal of Economics*, 119(1), 65–90.

Briec, W., K. Kerstens, and I. Van de Woestyne (2018), "Hypercongestion in Production Correspondences: An Empirical Exploration", *Applied Economics*, 50(27), 2938–56.

Byrnes, P. and R. Färe (1987), "Surface Mining of Coal: Efficiency of US Interior Mines", *Applied Economics*, 19(12), 1665–73.

Byrnes, P., R. Färe, S. Grosskopf, and S. Kraft (1987), "Technical Efficiency and Size: The Case of Illinois Grain Farms", *European Review of Agricultural Economics*, 14(4), 367–81.

Byrnes, P., R. Färe, S. Grosskopf, and C. Lovell (1988), "The Effect of Unions on Productivity: U.S. Surface Mining of Coal", *Management Science*, 34(9), 1037–53.

Çakmak, E. and O. Zaim (1992), "Türkye'de Tarim Kesiminde Etkinlik [Technical Efficiency of Turkish Agriculture]", *ODTÜ Gelişme Dergisi [METU Studies in Development]*, 19(3), 305–16.

Cooper, W., R. Thompson, and R. Thrall (1996), "Introduction: Extensions and New Developments in DEA", *Annals of Operations Research*, 66(1), 3–45.

Davutyan, N., M. Bilsel, and M. Tarcan (2014), "Migration, Risk-Adjusted Mortality, Varieties of Congestion and Patient Satisfaction in Turkish Provincial General Hospitals", *Data Envelopment Analysis Journal*, 1(2), 135–69.

De Borger, B., K. Kerstens, D. Prior, and I. Van de Woestyne (2012), "Static Efficiency Decompositions and Capacity Utilisation: Integrating Technical and Economic Capacity Notions", *Applied Economics*, 44(31), 4125–41.

Dervaux, B., G. Ferrier, H. Leleu, and V. Valdmanis (2004), "Comparing French and US Hospital Technologies: A Directional Input Distance Function Approach", *Applied Economics*, 36(10), 1065–81.

Dervaux, B., K. Kerstens, and P. Vanden Eeckaut (1998), "Radial and Nonradial Static Efficiency Decompositions: A Focus on Congestion Measurement", *Transportation Research: Part B: Methodological*, 32(5), 299–312.

Fan, Y., Q. Li, and A. Weersink (1996), "Semiparametric Estimation of Stochastic Production Frontier Models", *Journal of Business & Economic Statistics*, 14(4), 460–8.

Färe, R. (1975), "Efficiency and the Production Function", *Zeitschrift für Nationalökonomie*, 35(3–4), 317–24.

Färe, R. and S. Grosskopf (2000), "Decomposing Technical Efficiency with Care", *Management Science*, 46(1), 167–8.

Färe, R., S. Grosskopf, and J. Logan (1983a), "The Relative Efficiency of Illinois Electric Utilities", *Resources and Energy*, 5(4), 349–67.

Färe, R., S. Grosskopf, and J. Logan (1985a), "The Relative Performance of Publicly-owned and Privately-Owned Electric Utilities", *Journal of Public Economics*, 26(1), 89–106.

Färe, R., S. Grosskopf, and J. Logan (1987), "The Comparative Efficiency of Western Coal-Fired Steam-Electric Generating Plants: 1977-1979", *Engineering Costs and Production Economics*, 11(1), 21–30.

Färe, R., S. Grosskopf, J. Logan, and C. Lovell (1985b), "Measuring Efficiency in Production: With an Application to Electric Utilities", in *Managerial Issues in Productivity Analysis*, ed. A. Dogramaci and N. Adam, Boston: Kluwer, 185–214.

Färe, R., S. Grosskopf, and C. Lovell (1983b), "The Structure of Technical Efficiency", *Scandinavian Journal of Economics*, 85(2), 181–90.

Färe, R., S. Grosskopf, and C. Lovell (1985c), *The Measurement of Efficiency of Production*, Boston: Kluwer.

Färe, R., S. Grosskopf, and C. Lovell (1994), *Production Frontiers*, Cambridge: Cambridge University Press.

Färe, R., S. Grosskopf, and C. Pasurka (1989), "The Effect of Environmental Regulations on the Efficiency of Electric Utilities: 1969 versus 1975", *Applied Economics*, 21(2), 225–35.

Färe, R. and L. Svensson (1980), "Congestion of Production Factors", *Econometrica*, 48(7), 1745–53.

Farrell, M. (1957), "The Measurement of Productive Efficiency", *Journal of the Royal Statistical Society Series A: General*, 120(3), 253–81.

Field, K. (1990), "Production Efficiency of British Building Societies", *Applied Economics*, 22(3), 415–26.

Field, K. and A. Emrouznejad (2003), "Measuring the Performance of Neonatal Care Units in Scotland", *Journal of Medical Systems*, 27(4), 315–24.

Flegg, T. and D. Allen (2009a), "Congestion in the Chinese Automobile and Textile Industries Revisited", *Socio-Economic Planning Sciences*, 43(3), 177–91.

Flegg, T. and D. Allen (2009b), "Congestion in the New British Universities: A Further Analysis", *Journal of the Operations Research Society of Japan*, 52(2), 186–203.

Flegg, T., D. Allen, T. Thurlow, and K. Field (2004), "Measuring the Efficiency of British Universities: A Multi-Period Data Envelopment Analysis", *Education Economics*, 12(3), 231–49.

Førsund, F. and L. Hjalmarsson (1974), "On the Measurement of Productive Efficiency", *Swedish Journal of Economics*, 76(2), 141–54.

Frantz, R. (1988), *X-Efficiency: Theory, Evidence and Applications*, Boston: Kluwer.

Fukuyama, H. (1997), "Investigating Productive Efficiency and Productivity Change of Japanese Life Insurance Companies", *Pacific-Basin Finance Journal*, 5(4), 481–509.

Fukuyama, H. and W. Weber (1999), "The Efficiency and Productivity of Japanese Securities Firms, 1988-93", *Japan and the World Economy*, 11(1), 115–33.

Glass, J., D. McKillop, and G. O'Rourke (1997), "Productivity Growth in UK Accountancy Departments 1989-96", *Financial Accountability & Management*, 13(4), 313–30.

Glass, J. and D. McKillop (2000), "A Post Deregulation Analysis of the Sources of Productivity Growth in UK Building Societies", *Manchester School*, 68(3), 360–85.

Habibullah, M., M. Makmur, W. Azman-Saini, A. Radam, and H.-B. Ong (2005), "Bank Efficiency and the Efficient Market Hypothesis: The Case for Bank Stock Prices in KLSE", *Savings and Development*, 29(4), 363–90.

Hackman, S. (2008), *Production Economics: Integrating the Microeconomic and Engineering Perspectives*, Berlin: Springer.

Haskel, J. and A. Sanchis (2000), "A Bargaining Model of Farrell Inefficiency", *International Journal of Industrial Organization*, 18(4), 539–56.

Henderson, D. and R. Russell (2005), "Human Capital and Convergence: A Production-Frontier Approach", *International Economic Review*, 46(4), 1167–205.

Johnson, G. and W. Happ (1977), "Digital Simulation for Detecting Congestion in Hospital Facilities", in *Winter Simulation Conference Proceedings*, Institute of Electrical and Electronics Engineers (IEEE), 849–53.

Jones, J. (2001), *Laboratory Guide for Conducting Soil Tests and Plant Analysis*, Boca Raton: CRC Press.

Kao, C. (2010), "Congestion Measurement and Elimination under the Framework of Data Envelopment Analysis", *International Journal of Production Economics*, 123(2), 257–65.

Kneip, A., L. Simar, and P. Wilson (2016), "Testing Hypotheses in Nonparametric Models of Production", *Journal of Business & Economic Statistics*, 34(3), 435–56.

Kritikos, M., R. Markellos, and G. Prastacos (2010), "Corporate Real Estate Analysis: Evaluating Telecom Branch Efficiency in Greece", *Applied Economics*, 42(9), 1133–43.

Li, Q. (1996), "Nonparametric Testing of Closeness between Two Unknown Distribution Functions", *Econometric Reviews*, 15(1), 261–74.

Luenberger, D. (1995), *Microeconomic Theory*, Boston: McGraw-Hill.

McCallion, G., J. Glass, R. Jackson, C. Kerr, and D. McKillop (2000), "Investigating Productivity Change and Hospital Size: A Nonparametric Frontier Approach", *Applied Economics*, 32(2), 161–74.

McDonald, J. (1996), "A Problem with the Decomposition of Technical Inefficiency into Scale and Congestion Components", *Management Science*, 42(3), 473–4.

Mehdiloozad, M. and V. Podinovski (2018), "Nonparametric Production Technologies with Weakly Disposable Inputs", *European Journal of Operational Research*, 266(1), 247–58.

Mulumba, Z., L. Nalubanga, C. Nankanja, K. Manasseh, J. Månsson, and J. Hollén (2017), "Technical Efficiency Decomposed — The Case of Ugandan Referral Hospitals", *Central European Review of Economics and Management*, 1(4), 117–46.

Nasierowski, W. and F. Arcelus (2003), "On the Efficiency of National Innovation Systems", *Socio-Economic Planning Sciences*, 37(3), 215–34.

Ng, Y. and S.-K. Li (2009), "Efficiency and Productivity Growth in Chinese Universities during the Post-Reform Period", *China Economic Review*, 20(2), 183–92.

Odeck, J. (2006), "Congestion, Ownership, Region of Operation, and Scale: Their Impact on Bus Operator Performance in Norway", *Socio-Economic Planning Sciences*, 40(1), 52–69.

Olesen, O. and N. Petersen (2015), "Facet Analysis in Data Envelopment Analysis", in *Data Envelopment Analysis: A Handbook of Models and Methods*, ed. J. Zhu, New York: Springer, 145–90.

Olesen, O. and N. Petersen (2016), "Stochastic Data Envelopment Analysis–A Review", *European Journal of Operational Research*, 251(1), 2–21.

Paris, Q. (2008), "Law of the Minimum", in *Encyclopedia of Soil Science*, ed. W. Chesworth, New York: Springer, 431–7.

Patrick, J. (2011), "Access to Long-Term Care: The True Cause of Hospital Congestion?", *Production and Operations Management*, 20(3), 347–58.

Porembski, M., K. Breitenstein, and P. Alpar (2005), "Visualizing Efficiency and Reference Relations in Data Envelopment Analysis with an Application to the Branches of a German Bank", *Journal of Productivity Analysis*, 23(2), 203–21.

Puig-Junoy, J. (2000), "Partitioning Input Cost Efficiency into its Allocative and Technical Components: An Empirical DEA Application to Hospitals", *Socio-Economic Planning Sciences*, 34(3), 199–218.

Rajsic, P. and A. Weersink (2008), "Do Farmers Waste Fertilizer? A Comparison of *Ex Post* Optimal Nitrogen Rates and *Ex Ante* Recommendations by Model, Site and Year", *Agricultural Systems*, 97(1), 56–67.

Ray, S. (2004), *Data Envelopment Analysis: Theory and Techniques for Economics and Operations Research*, Cambridge: Cambridge University Press.

Russell, R. and W. Schworm (2009), "Axiomatic Foundations of Efficiency Measurement on Data-Generated Technologies", *Journal of Productivity Analysis*, 31(2), 77–86.

Russell, R. and W. Schworm (2011), "Properties of Inefficiency Indexes on ⟨Input, Output⟩ Space", *Journal of Productivity Analysis*, 36(2), 143–56.

Sauer, J. (2006), "Economic Theory and Econometric Practice: Parametric Efficiency Analysis", *Empirical Economics*, 31(4), 1061–87.

Scarf, H. (1994), "The Allocation of Resources in the Presence of Indivisibilities", *Journal of Economic Perspectives*, 8(4), 111–28.

Seitz, W. (1970), "The Measurement of Efficiency Relative to a Frontier Production Function", *American Journal of Agricultural Economics*, 52(4), 505–11.

Simões, P. and R. Marques (2011), "Performance and Congestion Analysis of the Portuguese Hospital Services", *Central European Journal of Operations Research*, 19(1), 39–63.

Syverson, C. (2011), "What Determines Productivity?", *Journal of Economic Literature*, 49(2), 326–65.

Thrall, R. (1996), "Duality, Classification and Slacks in DEA", *Annals of Operations Research*, 66(2), 109–38.

Wu, S., S. Devadoss, and Y. Lu (2003), "Estimation and Decomposition of Technical Efficiency for Sugarbeet Farms", *Applied Economics*, 35(4), 471–84.

Zeitsch, J. and D. Lawrence (1996), "Decomposing Economic Inefficiency in Base Load Power Plants", *Journal of Productivity Analysis*, 7(4), 359–78.

Zhengfei, G. and A. Oude Lansink (2003), "Input Disposability and Efficiency in Dutch Arable Farming", *Journal of Agricultural Economics*, 54(3), 467–78.