An Examination of Alternative Approaches to Measuring Congestion in British Universities

A.T. FLEGG & D.O. ALLEN[1]

ABSTRACT This paper examines three alternative methods of measuring congestion, from both theoretical and empirical perspectives. These methods include the conventional approach of Färe and Grosskopf, the alternative proposed by Cooper *et al*., and a new method developed by Tone and Sahoo. Each method is found to have merits and demerits. The properties of the different methods are examined using data for 45 British universities in the period 1980/81–1992/93. Despite conceptual differences, the results from Färe and Grosskopf's approach are found to be very similar indeed to those from Tone and Sahoo's approach. Contrary to expectations, Cooper's approach generally indicates less congestion than the other two procedures. In terms of the causes of congestion, excessive numbers of undergraduates are found to be the largest single cause of congestion in British universities during the period under review, although academic overstaffing is also identified as a major cause of congestion.

Introduction

In a recent paper in *Education Economics* (Flegg *et al*., 2004), we examine the impact on universities' efficiency of the rapid and unbalanced expansion in the period 1980/81–1992/93. This period is interesting because it was characterized by major changes in public funding, in student : staff ratios and in the management of universities. We find that around half of universities suffered from congestion in the sense that they could have produced a larger output by cutting down on one or more inputs. We argue that an excessive number of undergraduate students is the most likely cause of this congestion.

Along with most previous studies of congestion, the paper mentioned above follows the well-known procedure developed by Färe and Grosskopf. This has been criticized by Cooper *et al*., who recommend an alternative approach of their own. The issue of how to measure congestion has, in fact, engendered a heated debate in the *European Journal of Operational Research* and in *Socio-Economic Planning Sciences*. However, whilst the theoretical and measurement issues have been debated extensively, there is scant empirical evidence on whether the two approaches yield substantially different answers as regards the *measured* amount of congestion.

The primary aim of the present paper is, therefore, to see whether the two approaches produce noticeably different estimates of the amount of congestion in British universities in the period 1980/81–1992/93. In this regard, it is worth noting Färe and Grosskopf's observation that, of the two procedures, their approach would generally measure a *smaller* amount of congestion.

We also discuss a new approach to measuring congestion and scale economies, which has been put forward by Tone and Sahoo (2004), and present estimates of the amount of congestion indicated by their approach. Finally, we attempt to identify the extent to which the different inputs in our model contribute towards the observed amount of congestion.

We begin with a discussion of the theoretical properties of the different approaches and point out some advantages and disadvantages of each approach.

[1] Tony.Flegg@uwe.ac.uk and David.Allen@uwe.ac.uk. School of Economics, University of the West of England, Coldharbour Lane, Bristol BS16 1QY.

What is Congestion?

Cooper *et al.* (2001a, p. 62) define congestion in the following way:

Definition 1. *Input congestion occurs whenever increasing one or more inputs decreases some outputs without improving other inputs or outputs. Conversely, congestion occurs when decreasing some inputs increases some outputs without worsening other inputs or outputs.*

They go on to observe (*ibid.*, p. 63) that congestion can be regarded as a particularly severe form of technical inefficiency.

However, the above definition makes no reference to any limiting factor that might account for the congestion. A possible alternative definition might read as follows:

Definition 2. *Input congestion occurs whenever too much (little) of any input is employed, with all other inputs held constant, and this leads to a fall (rise) in output.*

This alternative definition takes explicit account of the hypothesis of diminishing marginal returns, with the added feature that congestion requires a fall (rise) in output.

Now consider the simple model $y = f(x_1, x_2)$, where y is some measure of educational output, x_1 is the number of academic staff and x_2 is the number of students. A necessary condition for congestion to exist is that one of these inputs has a negative marginal product. This will give rise to upward-sloping segments of the isoquants linking x_1 and x_2. The problem of congestion is the result of an excessive use of one or more inputs.

In the case of universities, it seems reasonable to assume that an *unbalanced* expansion could lead to congestion. For instance, in the period studied here (1980/81–1992/93), the number of students rose much more rapidly than the number of academic staff (see Figure 1). As a result, the marginal product of students might have become *negative* in some universities. The implication of this is that a reduction in the number of students, with all other inputs (staff, buildings, etc.) held constant, would raise the university's output in terms of research and degrees awarded, both undergraduate and postgraduate.

Measuring Congestion

The conventional way of measuring congestion was developed by Färe and Grosskopf, while Byrnes *et al.* (1984) and Färe *et al.* (1985a) were the first published applications. Cooper *et al.* (1996) then proposed an alternative procedure, which was refined and applied to Chinese data by Brockett *et al.* (1998) and Cooper *et al.* (2000). The merits and demerits of the two approaches have been debated most recently by Cherchye *et al.* (2001) and Cooper *et al.* (2001a,b). For ease of exposition, the two procedures are referred to hereafter as Färe's approach and Cooper's approach, with Färe and Cooper acting as representatives of the two schools of thought.

Färe's approach is an axiomatic one, which makes use of plausible assumptions about the nature of the productive technology (see Färe *et al.*, 1985b). It draws its inspiration from the theory of production and from the pioneering work of Farrell (1957). By contrast, Cooper's approach is more empirically based. It is grounded in the literature on Data Envelopment Analysis (DEA).

One of the main points of contention is how input slacks should be treated. Färe ignores such slacks on the basis that they can be disposed of at zero opportunity cost. Indeed, Färe and Grosskopf (2000, pp. 32–33) argue that, given positive input prices, non-zero slack is akin to *allocative* rather than *technical* inefficiency. By contrast, slacks are at the core of Cooper's slacks-based measure of congestion. Cooper *et al.* (2001a, p. 69) posit the following relationship (the notation has been simplified):

$$c_i = s_i^{-*} - \delta_i^* \tag{1}$$

where c_i is the amount of congestion associated with input i, s_i^{-*} is the total amount of slack in input i and δ_i^* is the amount of slack attributable to technical inefficiency. The measured amount of congestion is thus a residual derived from the DEA results.

Cooper *et al.* use the following apt example to illustrate the meaning of equation (1). Consider the difference between 'an excess number of workers exhibiting idle time but not otherwise interfering with production' and 'an excess of raw material inventory congesting a factory floor in a manner that interferes with production' (*ibid.*). The latter would represent congestion and would be captured by the variable c_i, whereas the former would represent technical inefficiency and would be measured by δ_i^*.

The differences between these two approaches are best illustrated by the use of examples.

Example 1 (see Figure 2)

Figure 2 shows seven DMUs, each producing an output of y = 2. This example, which assumes *constant returns to scale* and makes use of an *input-oriented approach*, is taken from Färe *et al.* (1985b, pp. 76–77). As regards DMUs C and D, there would be no dispute: both are clearly technically efficient. However, under Färe's approach, DMUs A and B would also be deemed to be efficient. Färe would disregard the fact that A and B have slack in x_2 of 2 units and 1 unit, respectively. By contrast, Cooper would treat these two DMUs as being only *weakly* efficient. Cooper regards non-zero slack as a form of technical inefficiency, whereas Färe argues that such slack can be ignored in an analysis of technical efficiency if it is *freely disposable*, i.e. where it can be disposed of at no opportunity cost.

The major differences between the two approaches arise with respect to DMUs E and F. Because F is on the isoquant for y = 2, Färe would regard this DMU as exhibiting no *pure* technical inefficiency (PTE = 1). However, it does appear to suffer from *congestion*. Its congestion efficiency (CE) score, as measured by the ratio OF′/OF, equals 0.8. Its technical efficiency (TE) score also equals 0.8 since TE is the product of PTE = 1 and CE = 0.8. According to Färe, congestion arises because of the difference between the upward-sloping isoquant segment DF, which is assumed to exhibit *weak* disposability, and the hypothetical horizontal dashed line emanating from D, which is assumed to exhibit *strong* (or free) disposability. By moving to point F′, F could attain TE = 1. This would be the end of the matter according to Färe. However, Cooper would then point to the non-zero slack of DF′ and say that this was indicative of technical inefficiency.

The case of E is more complicated because, according to Färe, this DMU suffers from both pure technical inefficiency and congestion. PTE is measured by the ratio OE′′/OE = 0.86 and CE by the ratio OE′/OE′′ = 0.93. Hence TE = 0.86 × 0.93 = 0.8. Färe would ignore the non-zero slack of DE′.

By contrast, Cooper would claim that there is no evidence that either E or F suffers from congestion! This is because all DMUs in Figure 2 produce the same output of y = 2. For congestion to occur, in his view, one must observe a fall in output if the input in question is increased or a rise in output if this input is reduced. For instance, if we move from E to F, raising the quantity of x_1 by one unit, there is no fall in y. Cooper's model, which divides any non-zero slack into technical and congestion components, would assign all of the non-zero slack of E and F to technical inefficiency (δ_i^* in equation (1) above).

In the context of this example, however, the criticisms of Färe's approach by Cooper *et al.* (2001a) are somewhat unfair. This is because, in an isoquant-type analysis, the DMUs are bound to have the same output and hence cannot possibly satisfy Cooper's definition of congestion! In a more realistic example, the DMUs would surely differ in terms of output. For example, suppose that we were to recast the present example slightly by raising the output of E from 2 to, say, 2.25 but leaving the output of all other DMUs constant at 2. If we now moved from E to F, the rise in x_1 from 3 to 4 would be accompanied by a *fall* in output from 2.25 to 2. Clearly, this would constitute 'congestion' in the sense of Definition 1 above.

What is more, even if all DMUs had y = 2, we could still validly argue that E and F suffered from congestion in input x_1. This is because, along segment DF, the marginal product of x_1 must be negative. Output stays constant along DF because the rise due to increased use of the non-congested input x_2 exactly offsets the fall due to increased use of the congested input x_1.

Example 2 (see Figure 3)

Figure 3 shows six DMUs. This example, which again makes use of an *input-oriented approach*, is taken from Cooper *et al.* (2001a). Whereas R produces an output of y = 10, the remaining DMUs all produce y = 1. *Variable returns to scale* are now assumed. The figure takes the form of a pyramid with its pinnacle at R. How would Färe evaluate these DMUs? A, B and R are clearly fully efficient. However, C, G and D would be deemed to be inefficient, with *all* of the inefficiency ascribed to the pure technical category. This, of course, would indicate an absence of congestion. This finding can be explained by the fact that the projections onto the efficiency frontier occur along segment BA, at points C′, G′ and D′. These three DMUs have TE = PTE = 0.4 and CE = 1.

Cooper would dispute the finding of no congestion in the case of G, arguing that there is, in fact, compelling evidence of its existence. For instance, suppose that we went from G to R. The inputs of both factors would fall by 2.5 units, yet there would be a tenfold rise in output!

However, as Färe and Grosskopf (2000a, p. 32) themselves point out, a segment like CD on the unit isoquant would be ruled out of order by their axiom of weak disposability. In their world, isoquants may not join up in this 'circular' fashion. Weak disposability means that a proportionate increase in both x_1 and x_2 cannot decrease output. This rules out the possibility that both factors might have negative marginal products, which is a necessary condition for a downward-sloping segment such as CD to occur. If we really did have a situation where both MP_1 and MP_2 were negative, then this would surely be a case of congestion! The case of G highlights a possible shortcoming of Färe's approach. Clearly, any DMU situated in between C and D would be in a similar situation.

It is worth considering what congestion might mean in the case of G. Cooper *et al.* (2001a,b) do not consider this issue, although they criticize Färe's approach on the grounds of its alleged adherence to the law of variable proportions. Cooper *et al.* (2001a, Table 4) define the region CDR in terms of the equation $y = 28 - 1.8x_1 - 1.8x_2$, which entails that *both*

marginal products must be negative. For this to make economic sense in terms of the law of variable proportions, there would need to be some latent factor that was being held constant. Alternatively, but less plausibly, one might argue that diseconomies of scale had become so severe that equiproportionate increases in both factors were causing output to fall. Cherchye *et al.* (2001, p. 77) note that this second possibility would be ruled out, in the case of Färe's approach, by the axiom of weak disposability.

The polar cases of C and D are interesting too because we must have $MP_1 > 0$, $MP_2 < 0$ along segment BC but $MP_1 < 0$, $MP_2 > 0$ along segment AD. The fact that one of the inputs has a negative marginal product in each case corresponds to an intuitive notion of congestion, yet Färe's approach does not validate this notion! In fact, his approach only signals the existence of congestion where the relevant upward-sloping segment of the isoquant is either relatively steep or relatively flat. To show this, let us move the positions of DMUs D and C, in turn.

If we move D to position D* in Figure 3, then TE = PTE = CE = 1. There is thus no congestion. This is also true if we move D to position D**, although TE = PTE = 0.5 at this point. However, congestion occurs in between D* and D** and increases as we move closer to the latter point. A similar analysis can be applied to C. In fact, any point in between D* and D** or in between C* and C** will have $0.5 < CE < 1$.

The above is not a very plausible outcome. Since the gradient of the isoquant equals $-MP_1/MP_2$, any isoquant segment lying in between AD* and AD** must have a relatively small (negative) value for MP_1 but a relatively large (positive) value for MP_2. Similarly, any isoquant segment lying in between BC* and BC** must have a relatively small (negative) value for MP_2 but a relatively large (positive) value for MP_1. Thus it would appear that Färe's approach tends to identify congestion when the factor in question has a marginal product that is only marginally negative but ignores it when the marginal product is highly negative! This seems counterintuitive.

Given these problems with Färe's approach, we might ask whether Cooper's approach would fare any better. Cooper *et al.* (2001a) do not mention the possibility of using the input-oriented variant of their method, so it is worth noting that this would yield the same outcome as Färe's approach with respect to DMUs C, G and D, i.e. no congestion. The reason is that non-zero input slacks are necessary (but not sufficient) for congestion to be identified and, in this instance, both methods would produce zero slacks in the first stage.

From the above discussion, it seems clear that input-oriented models are best avoided when attempting to identify and measure congestion. Therefore, in the next example, we will examine the use of output-oriented models.

Example 3 (see Figure 4)

Figure 4 shows the six DMUs from the previous example plus two more, P and Q. This example, which employs an *output-oriented approach*, is adapted from Brockett *et al.* (1998). *Variable returns to scale* are again assumed.

In the output-oriented variant of Färe's approach (as used, for example, in Byrnes *et al.*, 1984, and Färe *et al.*, 1985a), the congestion score, C_F, is calculated using the following ratio:

$$C_F = \phi^*/\beta^* \qquad (2)$$

where ϕ^* and β^* are the optimal values derived from the stage 1 and 2 models, respectively. Note that $C_F \geq 1$, with $C_F = 1$ indicating an absence of congestion. The stage 1 and stage 2 models for a particular DMU k are shown below (cf. Cooper *et al.*, 2000, pp. 3–6):

$$\phi^* = \max \phi \tag{3a}$$

$$\text{s.t.} \quad \sum_j \lambda_j x_{ij} \leq x_{ik} \qquad i = 1, 2, \ldots, m \tag{3b}$$

$$\sum_j \lambda_j y_{rj} \geq \phi y_{rk} \qquad r = 1, 2, \ldots, s \tag{3c}$$

$$\sum_j \lambda_j = 1 \tag{3d}$$

$$\lambda_j \geq 0 \qquad j = 1, 2, \ldots, n \tag{3e}$$

$$\beta^* = \max \beta \tag{4a}$$

$$\text{s.t.} \quad \sum_j \lambda_j x_{ij} = \gamma x_{ik} \qquad i = 1, 2, \ldots, m \tag{4b}$$

$$\sum_j \lambda_j y_{rj} \geq \beta y_{rk} \qquad r = 1, 2, \ldots, s \tag{4c}$$

$$\sum_j \lambda_j = 1 \tag{4d}$$

$$\lambda_j \geq 0 \qquad j = 1, 2, \ldots, n \tag{4e}$$

$$0 \leq \gamma \leq 1 \tag{4f}$$

Cooper *et al.* (2000, pp. 12–13) use the above output-oriented model to examine the situation facing DMU G. In doing so, they find $\phi^* = 10$, $\lambda_R^* = 1$, $s_1^{-*} = s_2^{-*} = 2.5$ in stage 1. In stage 2, they find $\beta^* = 10$, $\lambda_R^* = 1$ and $\gamma^* = 5/7.5 = ⅔$. These results yield $C_F = \phi^*/\beta^* = 1$, so no congestion is indicated. The reason for this outcome is that the same DMU, viz R, is being used to evaluate G in both stages. In stage 2, G's inputs are scaled down by $\gamma^* = ⅔$ to reach the levels attained by R.

Brockett *et al.* (1998) also examine a point like Q in Figure 4. Although this produces $C_F > 1$ and thus suggests the existence of congestion, making use of Färe's output-oriented model in this way would not be consistent with his methodology.[2] As noted earlier, the axiom of weak disposability means that *any* DMU located in between C and D would be free of congestion, with its inefficiency being deemed to be purely technical in nature.

However, neither Cooper *et al.* (2000) nor Brockett *et al.* (1998) examine the case of DMUs C and D, so let us now examine whether they suffer from congestion. As regards D, stage 1 yields $\phi^* = 10$, $\lambda_R^* = 1$, $s_1^{-*} = 5$, $s_2^{-*} = 0$. For stage 2, $\beta^* = 4.375$, $\lambda_R^* = 0.375$, $\lambda_A^* = 0.625$ and $\gamma^* = 0.5$. In this instance, $C_F = 10/4.375 \approx 2.286$, so the output-oriented version of Färe's procedure shows that D does indeed suffer from congestion. This is in contrast to the input-oriented version, which suggested that D was free from congestion. In stage 2, D is projected

[2] It should also be noted that Brockett *et al.* project their DMU Q in stage 2 to a point like Q′ on the congested ridge line CR (where y and x_2 vary inversely) rather than to a point like T on the congestion-free ridge line BR. However, β^* is not maximized at Q′. In addition, the co-ordinates they give for Q are not consistent with this DMU being on the frontier.

to D′, a point on the congestion-free ridge line AR, by scaling its inputs down by $\gamma^* = 0.5$. C_F is then computed as the ratio of the outputs at R and D′. In like fashion, C is projected on to the congestion-free ridge line BR. $C_F = 0/4.375 \approx 2.286$ in this case, so that C and D suffer from an identical amount of congestion.

It is of some interest to establish what happens to the congestion scores along the segments BC and AD of the efficiency frontier. Although Brockett *et al.* (1998) examine a DMU akin to P in Figure 4, the co-ordinates they give are not consistent with this DMU being on the frontier. Hence their analysis was reworked as follows.

Stage 1 for P yields $\phi^* = 5.5$, $\lambda_B^* = \lambda_R^* = 0.5$, $s_1^{-*} = 0$, $s_2^{-*} = 2.5$. P is projected to point S in Figure 4, where y = 5.5. In stage 2, we get $\beta^* = 3.25$, $\lambda_R^* = 0.25$, $\lambda_B^* = 0.75$ and $\gamma^* = ⅔$. The projection here is to point P′, where y = 3.25. Hence $C_F = 5.5/3.25 \approx 1.692$. Thus, from the diagram, it is evident that the value of C_F will rise monotonically from unity at B to reach a maximum of 2.286 at C. Line AD will exhibit the same property.

If we accept – as the present authors do – that all points lying on the segments BC, CD and AD of the frontier in Figure 4 are congested (since the marginal product of x_1 or x_2 or both is negative), then the output-oriented version of Färe's procedure is clearly more successful than the input-oriented version at identifying the congestion that exists.[3] However, any DMU located along CD in between C and D would be deemed to be suffering from pure technical inefficiency rather than congestion. From our perspective, this is a serious shortcoming of Färe's procedure. Hence, in the next example, we shall be looking at Cooper's method with this aspect particularly in mind.

Example 4 (see Figure 5)

Before examining Figure 5, which is adapted from Brockett *et al.* (1998), we need to define Cooper's measure of congestion, denoted here by C_C. The first step is to rewrite equation (1) as follows:

$$c_i/x_i = s_i^{-*}/x_i - \delta_i^*/x_i \tag{5}$$

where c_i/x_i is the proportion of congestion in input i, s_i^{-*}/x_i is the proportion of slack in input i and δ_i^*/x_i is the proportion of technical inefficiency in input i. The second step is to take arithmetic means over all m inputs to get:

$$C_C = \overline{s/x} - \overline{\delta/x} \tag{6}$$

Hence C_C measures the average proportion of congestion in the inputs used by a particular DMU. It has the property $0 \le C_C \le 1$. See Cooper *et al.* (2001a, p. 73).

The first stage of Cooper's procedure makes use of the output-oriented version of the BCC model. This, in turn, involves two steps. In the first step, model (3) is employed to obtain the value of ϕ^* for each DMU, while the second step involves maximizing the sum of the slacks, conditional on this value of ϕ^*. To illustrate, consider DMU E in Figure 5.

[3] $MP_1 < 0$ for $x_1 > 5$ and $MP_2 < 0$ for $x_2 > 5$.

Figure 5 reveals that there are two possible referent DMUs available for evaluating DMU E, viz B and C. Both would yield $\phi^* = 2$, yet B is the DMU that would maximize the slack in input x (giving $s_x^- = 3$ versus only 2 for C). Hence B is the DMU picked out in stage 1.

In stage 2 of Cooper's procedure, the slacks are again maximized but subject, in this case, to the projected output remaining constant. Hence, in Figure 5, we would move along the BCC frontier from B to C, holding output constant at y = 2. This process would yield $\delta_x^* = 1$.

Hence, in the case of DMU E, the three units of slack in input x obtained from the BCC model would be divided into two units of congestion and one unit of technical inefficiency. In terms of equation (6), we would have $\overline{s/x} = 3/5$ and $\overline{\delta/x} = 1/5$, giving $C_C = 0.4$. As regards the other DMUs, this method would generate $C_C = 0.25$ for D and F. G and H would be free from congestion, as would C. D would have $\phi^* = 2/1.5 = 1\frac{1}{3}$, whereas F, G and H would have $\phi^* = 2$. The figure also illustrates the point that the presence of non-zero slack is necessary but not sufficient for congestion to occur. It is worth noting, finally, that the input-oriented version of Cooper's approach would have shown no congestion for DMU E, thereby again illustrating the disadvantages of this orientation when measuring congestion (the projection would have been to point E′ in Figure 5).

In real data sets, horizontal segments such as BC in Figure 5 are rare and, in our own data set of 45 universities over 13 years, all universities are BCC efficient. If the BCC frontier does not have any DMUs like C, then the amount of congestion for each input equals the BCC slack for this input. This greatly simplifies the work needed to compute C_C, since stage 2 of Cooper's procedure can be skipped.

Let us now return to Figure 4 to see how Cooper's approach would evaluate the DMUs shown there. In the case of G, we get $C_C = \frac{1}{2}\{(2.5/7.5) + (2.5/7.5)\} = \frac{1}{3}$. $C_C = 0.25$ for C and D. For Q, we get $C_C = \frac{1}{2}\{(1/6) + (4/9)\} \approx 0.306$. These results show a modest *rise* in the amount of congestion as we approach G from either side, which is more plausible than the outcome from Färe's output-oriented model. As regards segment BC, the value of C_C rises monotonically from zero at B to $\frac{1}{2}\{0 + (2.5/7.5)\} \approx 0.167$ at P, reaching a maximum of 0.25 at C. The same thing happens along segment AD. The same property was found in the case of Färe's output-oriented model, although the rise in C_C is more modest than that in C_F.

It seems fair to conclude from the examples and procedures considered thus far that Cooper's output-oriented measure of congestion generates the most satisfactory results. However, there are some other considerations that need to be borne in mind.

Pros and Cons of the Two Approaches

The most attractive feature of Färe's approach is that it is possible to decompose overall technical efficiency (TE) in a straightforward way into pure technical efficiency (PTE), scale efficiency (SE) and congestion efficiency (CE), using the identity:

$$TE \equiv PTE \times SE \times CE \qquad (7)$$

Moreover, these measures can readily be incorporated into a *Malmquist analysis* to examine trends in efficiency over time (see Färe *et al.*, 1992, 1994; Flegg *et al.*, 2004). In terms of software, one can use *OnFront* (www.emq.com) to carry out the necessary calculations. On the other hand, we would argue that Färe's approach has a number of shortcomings:

- It rules out *a priori* certain aspects of production that do not fit into its theoretical framework, e.g. where both factors in a two-input model have negative marginal products.

- Only certain instances of negative marginal productivity are deemed to constitute congestion. What is more, our earlier discussion suggested that these cases were not the most plausible ones.

- The theoretical constructs underlying this approach are complex, as is the associated terminology. This makes it difficult to interpret the results.

- DMUs on the frontier may be weakly rather than strongly efficient.

However, in defending Färe's approach, Cherchye *et al.* (2001, pp. 77–78) point out that the original purpose of this procedure was not to measure the amount of congestion *per se* but instead to measure the impact, if any, of congestion on the overall efficiency of a particular DMU. This is a valid and important point, which can explain why Färe and his associates would insist that DMU G in Figures 3 and 4 does not exhibit congestion. Nevertheless, many researchers – including the present authors – have used Färe's methodology to identify and measure congestion, so it is also important to establish whether it performs this additional task correctly.

Of the two variants of Färe's approach, the examples discussed earlier suggested that the output-oriented version was the most satisfactory one to use for identifying and measuring congestion. However, we need to recognize that Färe would be loath to use an output-oriented model with variable returns to scale to examine the efficiency of the DMUs shown in Figure 4. This is because they produce a single output and because of the problems associated with distinguishing between scale inefficiency and congestion.[4] We also need to bear in mind the hazards of generalizing from a particular numerical example about the relative performance of different approaches (see Cherchye *et al.*, 2001, p. 76).

The most attractive feature of Cooper's approach is that it makes use of concepts that can readily be identified and measured in a set of data. On the basis of the examples considered here, the output-oriented variant of his approach appears to work well and to produce plausible results. What is more, his measure of congestion, C_C, is easy to understand and one can immediately see which factors are causing the problem and to what extent. By contrast, this information is more difficult to obtain from Färe's procedure (see Cooper *et al.*, 2000, pp. 6–7). However, a demerit of Cooper's non-radial methodology is that a straightforward decomposition of overall technical efficiency cannot be carried out. In addition, it is not entirely clear what aspects of the data Cooper's formula is trying to capture: is it negative marginal productivity or severe scale diseconomies or both?

To compute C_C, one needs to run a BCC output-oriented model to obtain the input slacks that underlie this measure, and then carry out some further calculations to work out $\overline{s/x}$ in equation (6) for each DMU. We used the *DEA-Solver Pro* software (www.saitech-inc.com) to generate the slacks and Excel to perform the calculations.

Whilst there are clear and fundamental conceptual differences between the two approaches, it is not yet clear whether they would produce very different results in reality, although we

[4] We are indebted to Pontus Roos of EMQ for pointing this out. He referred us to Färe and Grosskopf (2000b). *OnFront* uses constant returns to scale as the reference technology when decomposing TE into PTE, SE and CE.

should note the observation by Färe and Grosskopf (2000a, pp. 32–33) that their approach would generally measure a smaller amount of congestion. This contention is supported by the findings of Cooper *et al.* (2000), who examined data for three Chinese industries (textiles, chemicals and metallurgy) over the period 1966–88 and obtained noticeably larger amounts of congestion when their own method was employed. In the present paper, we aim to add to the scant empirical evidence on this topic.

A New Approach to Measuring Congestion

Tone and Sahoo (2004) have proposed a new unified approach to measuring congestion and scale economies. This has several attractive features. The first is that, unlike Färe's method, negative marginal productivity always signals congestion. This is as it should be. Secondly, the analysis can easily be done using *DEA-Solver Pro*. Thirdly, the output is comprehensive and easily understood. For simplicity, this procedure is referred to hereafter as Tone's approach.

Tone uses an output orientation. In fact, his approach is similar to Cooper's output-oriented method inasmuch as a BCC output-oriented model is used in the first stage. However, it differs in the second stage in its use of a slacks-based model. To explain this approach, let us return to the example in Figure 4.

Like Cooper, Tone would find A, B and R to be BCC efficient and hence not congested. The remaining DMUs (apart from P) would have a congestion score, C_T, of 10, reflecting the fact that R is producing ten times as much output as any of them. A more interesting bit of output from *DEA-Solver* is the figure for the *scale diseconomy*, ρ. For example, in the case of C, this is calculated as:

$$\rho = \frac{\%\text{ change in y}}{\%\text{ change in x}_2} = \frac{+900\%}{-50\%} = -18 \qquad (8)$$

Using the same method, we also get ρ = –18 for D. In the case of G, the average percentage change in inputs is –33⅓%, so that ρ = –27. These results suggest that congestion is equally serious for C and D but more serious for G. This finding is consistent with the outcome from Cooper's approach, where C_C = ⅓ for G but 0.25 for C and D. In Tone's terminology, we would describe G as being *strongly* congested (because both inputs are congested) but C and D as being *weakly* congested (because only one input is congested).

Findings Using the Different Approaches

In Flegg *et al.* (2004), we examined annual data for 45 British universities in the period 1980/81–1992/93. Our model included three outputs and four inputs. The outputs were:

- income from research and consultancy;
- the number of undergraduate degrees awarded, adjusted for quality;[5]
- the number of postgraduate degrees awarded.

[5] To adjust for quality, the number of undergraduate degrees awarded was multiplied by the proportion of first-class degrees, giving the *number* of first-class degrees as the output variable.

The inputs comprised:

- the number of academic and academic-related staff;
- the number of undergraduate students;
- the number of postgraduate students;
- aggregate departmental expenditure.[6]

Based on this DEA model, the output-oriented variant of Färe's approach was used to compute a congestion efficiency (CE) score for each university. A weighted geometric mean (WGM) was then calculated for each year, using the number of students in each university as a weight. We found that the WGM score, $\overline{CE}$, rose from 0.942 in 1980/81 to 0.967 in 1992/93. Within this period, $\overline{CE}$ rose steadily between 1984/85 and 1988/89 but fluctuated markedly thereafter. Notwithstanding these fluctuations in $\overline{CE}$, however, the number of universities exhibiting congestion remained high throughout the study period (the range was from 19 to 26).

The aim now is to see whether it makes much difference *how* congestion is measured. We also hope to shed some more light on the factors underlying the congestion observed during the study period. Our findings are presented in Tables 1 to 4 and the accompanying figures.

Table 1 shows the annual mean scores for the three approaches, calculated in different ways. The top panel shows the results from Färe's output-oriented approach. For consistency with the other methods, the reciprocal of each university's CE score was taken before computing annual means. For example, the WGM of 1.034 in Table 1 for 1992/93 corresponds to $\overline{CE} = 0.967$. One can see that the use of geometric rather than arithmetic means invariably yields lower values (e.g. compare the columns headed WAM and WGM). By contrast, weighting the raw congestion scores by the number of students has the effect of raising the annual means from 1981/82 onwards, indicating greater congestion.

The middle panel of Table 1 shows the results from Tone's approach. The impact of weighting and using geometric means is very similar to that shown in the top panel. This is also the case when we examine the results from Cooper's approach, shown in the bottom panel. It should be noted that, again for consistency, this panel shows the mean values of $1 + C_C$.

Table 2 compares the annual mean scores for the three approaches. Looking at the results for the WGM, one can see that, for nine of the thirteen years, Färe's approach indicates the highest amount of congestion, followed by Tone's approach and then Cooper's. This pattern is very clear up to 1987/88. Thereafter, the rankings are more changeable. The three methods are unanimous that 1984/85 was the most congested academic year in British universities. However, there is some disagreement over the least congested year, with Färe and Tone selecting 1992/93 but Cooper opting for 1990/91. The table also reveals that the maximum difference across methods occurs in 1982/83 and the minimum in 1990/91.

Figure 6a shows that the mean scores obtained from Färe's method are invariably a little higher than those from Tone's method. There is clearly an extremely strong correlation between these two measures of congestion. What is more, this very close relationship is not disturbed by the introduction of weighting in Figure 6b.

[6] This is total departmental recurrent expenditure *other than* that on academic and academic-related staff *plus* departmental equipment expenditure, summed over all departments in a given university.

The close relationship between the measures associated with Färe and Tone can be explained by similarities in the methods of calculation. In stage 1 of Tone's procedure, an output-oriented BCC model is employed. Whilst this model differs from that used in the first stage of Färe's method (see equations (3a) to 3(e)), the absence of any weakly efficient frontier universities in the present data set – universities akin to DMU C in Figure 5 – means that the first stages of the two methods invariably yield identical results. However, differences do arise in the second stages. This is because Färe employs a radial (i.e. proportional) projection to eliminate congestion whereas Tone uses an output-oriented version of the slacks-based model, which is a non-radial approach.

An examination of the values of C_F and C_T for individual universities confirmed the very strong correlation between these two measures. Of the $13 \times 45 = 585$ cases compared, $C_F = C_T = 1$ in 285 cases (48.7%) and $C_F = C_T > 1$ in 221 cases (37.8%). Of the remaining 79 cases (13.5%), 78 had $C_F > C_T$ and one had $C_F > 1$ but $C_T = 1$. It is worth noting that differences arose only in cases where PTE < 1. Since such cases were comparatively rare (an annual average of 11), this severely constrained the number of possible instances of $C_F \neq C_T$. In fact, the average number of cases of $C_F \neq C_T$ was only 6 per year.

Even though the differences between the values of C_F and C_T were fairly small, we were nonetheless surprised to find that invariably $C_F \geq C_T$. We had expected the opposite because, with Tone's procedure, negative marginal productivity entails congestion, whereas this is not necessarily so under Färe's approach.

It is interesting that Cooper's procedure indicates a substantially *lower* amount of congestion than the other two approaches in the period up to 1986/87. This is not what we expected. However, from 1987/88 onwards, his measure appears to converge with those of Färe and Tone, before moving higher at the end of the study period. When weighted means are used, Cooper's approach clearly signals higher congestion in the last two years. Looking at the WGM graphs, one can see that Cooper's method misses out on the rise in congestion shown by the other two methods in 1982/83 but mirrors the rise in 1984/85. It is evident that the weighting has more impact on Cooper's measure than on the other two.

How can the relatively low amounts of congestion indicated by Cooper's approach be explained? One possibility is purely technical: $1 + C_C$ has a range of [1, 2] whereas C_F and C_T have a minimum of one but no finite maximum. This problem could perhaps be circumvented by comparing C_C with $1 - CE_k$, where CE_k is the congestion efficiency score of university k; $0 \leq CE_k \leq 1$. All measures would then be constrained to the interval [0, 1].

Figures 7a and 7b show the effects of using a constrained interval for Färe's measure of congestion. As expected, the gap between Cooper's measure, C_C, and the revised version of Färe's measure has narrowed considerably, most noticeably in the peak year of 1984/85. The same thing happens when the measures are weighted, but it is worth noting that Cooper's measure still indicates more congestion in the last two years.

Perhaps we should not be surprised that Cooper's procedure does not yield an unambiguously higher or, indeed, lower measured amount of congestion than Färe's approach. This is because Cooper is measuring the average proportion of congestion in a given university's inputs, whereas Färe (and Tone) are measuring the potential gain in output from eliminating this congestion. These are related but different aspects of the same phenomenon, so that the expected relative size of the measures is hard to determine *a priori*.

Also, our expectation that Cooper's procedure would indicate more congestion than Färe's approach was based partly on the empirical findings of Cooper *et al.* (2000), who examined

data for three Chinese industries (textiles, chemicals and metallurgy) over the period 1966–88 and obtained noticeably larger amounts of congestion when their own method was employed. However, their study and ours are not directly comparable. This is because we have used panel data for 45 universities over thirteen years, whereas they used annual time-series data for each industry, with each year being treated as a separate 'DMU'. Another important difference is that we have three outputs in our model, whereas they had only one. These differences may well explain the different outcomes.[7]

The first column of Table 3 shows the annual arithmetic mean values of ρ, Tone's *scale diseconomies* parameter, based on data for all 45 universities. The table then shows the effect of excluding non-congested universities. Given a 1% decrease in congested inputs, the results indicate a potential rise in output of 5.9% on average in 1982/83 but only 1.7% in 1989/90. This suggests that congestion was much more serious in 1982/83. It should be noted that only congested inputs are included in the calculation of ρ. Likewise, only those outputs affected by congestion are considered, i.e. those where non-zero slack indicates a potential rise in output. Hence ρ does not measure the ratio of the overall percentage changes in inputs and outputs.

Whereas ρ suggests that congestion was most serious in 1982/83 but least serious in 1989/90, C_T picks out 1984/85 as the year with the most congestion and 1992/93 as the year with the least (see Table 2). At first sight, this disagreement is somewhat surprising. However, the differences in the values of $\overline{\rho}$ for 1989/90 and 1992/93 are negligible. As regards 1982/83, this year includes some atypically large values, with six universities having $|\rho| > 10$, which partly explains the relatively large value of $\overline{\rho}$. An examination of the data revealed that ρ was much more prone than C_T to fluctuate from year to year.[8] For example, for Aberdeen, $|\rho|$ rose dramatically from zero in 1982/83 to 31.7 in 1984/85, whereas C_T rose more gently from zero to 0.134. Also, in the case of Reading, $|\rho|$ fell sharply from 20 in 1982/83 to zero in 1984/85, whereas C_T fell less noticeably from 0.027 to zero. In both cases, the congestion was associated with a large shortfall in the number of first-class degrees awarded.

It is evident that C_T and ρ are unlikely to yield the same ranking of years in terms of the overall amount of congestion. Nonetheless, each measure provides some very useful but different information, so they should be seen as complementary.

Tables 4a and 4b take a closer look at the results from Cooper's method. Table 4a shows how C_C was calculated in each year (using unweighted arithmetic means), while Table 4b shows the contribution of each input to the mean value of C_C. This decomposition reveals several striking features. We can see that, in all years, excessive numbers of undergraduates were the largest single cause of congestion in British universities, accounting for between 39% and 53% of the value of Cooper's congestion score. However, it is also apparent that academic overstaffing was also a major cause of congestion! Indeed, in 1988/89, academic staff and undergraduates accounted for almost the same proportion of C_C. The table indicates that postgraduates had a substantially smaller impact than undergraduates in terms of causing congestion, although the gap between their respective contributions was much larger at the outset of the study period than at the end. Finally, we can see that excessive departmental expenditure played a negligible role in producing congestion.

---

[7] Cooper *et al.* (2000) compared C_C with $C_F - 1$. However, $C_F - 1$ has no finite upper limit and thus suffers from the same problem as our measure, C_F. We opted for C_F rather than $C_F - 1$ so we could take geometric means.

[8] For n = 45, the coefficient of variation for C_T ranged from 0.047 to 0.175 over the study period.

The finding regarding academic overstaffing is puzzling. What it suggests is that a reduction in the number of academic staff, other things being equal, could have *raised* the output of congested universities in terms of earnings from research and consultancy, first-class degrees awarded and postgraduate degrees obtained. One possible explanation is that overstaffing caused congestion of facilities such as libraries, office accommodation, etc. and this, in turn, caused a fall in output. This would be relevant if the frontier universities were generally better endowed than the congested universities. It is also possible that the 'surplus' staff in the congested universities were generally less qualified and experienced than their counterparts in the frontier universities. This might have reduced the average productivity of staff in the congested universities, although it is unlikely to have resulted in a negative marginal product. Unfortunately, we were unable to control for non-homogeneity of staff or students.

Conclusion

This paper has examined three alternative methods of measuring congestion, from both theoretical and empirical perspectives. The theoretical discussion suggested that an output-oriented approach was preferable to an input-oriented one when attempting to identify congestion and measure its extent. What is more, one could argue that an objective of maximizing output from given resources is much closer to what British universities are likely to be aiming for than the alternative of minimizing the resources used to produce a given output.

The theoretical discussion identified some shortcomings of Färe and Grosskopf's procedure for measuring congestion. Nonetheless, this conventional approach is still useful if one's aim is to assess the impact of congestion on the overall technical efficiency of a given university. What is more, this overall efficiency score can easily be decomposed into scale, congestion and pure technical components. This point is particularly germane when the DEA is being used in conjunction with a Malmquist analysis of the trends in efficiency over time.

In general, the method developed by Cooper and his associates appears to be superior to Färe and Grosskopf's procedure in terms of its ability to measure the extent of congestion and to shed light on its underlying causes. The new unified approach to measuring congestion and scale economies proposed by Tone and Sahoo (2004) also has several attractive theoretical and practical advantages. One of most important of these is the fact that, unlike Färe and Grosskopf's method, negative marginal productivity always signals congestion. Tone and Sahoo's method is also well supported by the *DEA-Solver Pro* software. However, a demerit of their approach is that their measure of congestion is not linked in a straightforward way to the other components of overall technical efficiency. The same problem arises with respect to the measure proposed by Cooper *et al.*

The three methods performed equally well in terms of identifying which universities were congested and which were not. However, there were differences in the amounts of congestion indicated by the different methods, although there was a high degree of similarity between the scores obtained from Färe and Grosskopf's method and Tone and Sahoo's method. Both methods showed a reduction in the amount of congestion in British universities in the period 1980/81 to 1992/93. This is a remarkable achievement, considering the rapid expansion in the number of students during this period, especially from 1988/89 onwards (see Figure 1).

Contrary to expectations, Cooper's procedure generally indicated *less* congestion than the other two methods, although this was not true in the last two years. For the study period as a whole, Cooper's procedure indicated a small rise in congestion.

The results revealed that, in all years, excessive numbers of undergraduates were the largest single cause of congestion in British universities during the period 1980/81–1992/93, accounting for between 39% and 53% of the value of Cooper's congestion score, C_C. However, it was apparent that academic overstaffing was also a major cause of congestion! Indeed, in 1988/89, academic staff and undergraduates accounted for almost the same proportion of C_C. The results also indicated that postgraduates had a substantially smaller role than undergraduates in causing congestion, although the gap between their respective contributions was much larger at the outset of the study period than at the end.

The finding regarding academic overstaffing is puzzling. Although some tentative suggestions were made regarding possible explanations, this facet of the results clearly warrants further investigation. In particular, we intend to explore whether this finding is a genuine one or merely an artefact of the specific inputs and outputs used in the model (especially the use of first-class degrees as an output variable). We also intend to carry out some statistical testing in an effort to see whether the findings regarding congestion are statistically significant.

References

Brockett, P.L., Cooper, W.W., Shin, H.C. & Wang, Y. (1998) Inefficiency and congestion in Chinese production before and after the 1978 economic reforms, *Socio-Economic Planning Sciences*, 32, pp. 1–20.

Byrnes, P., Färe, R. & Grosskopf, S. (1984) Measuring productive efficiency: an application to Illinois strip mines, *Management Science*, 30, pp. 671–681.

Cherchye, L., Kuosmanen, T. & Post, T. (2001) Alternative treatments of congestion in DEA: a rejoinder to Cooper, Gu, and Li, *European Journal of Operational Research*, 132, pp. 75–80.

Cooper, W.W., Gu, B. & Li, S. (2001a) Comparisons and evaluations of alternative approaches to the treatment of congestion in DEA, *European Journal of Operational Research*, 132, pp. 62–74.

Cooper, W.W., Gu, B. & Li, S. (2001b) Note: Alternative treatments of congestion in DEA – a response to the Cherchye, Kuosmanen and Post critique, *European Journal of Operational Research*, 132, pp. 81–87.

Cooper, W.W., Seiford, L.M. & Zhu, J. (2000) A unified additive model approach for evaluating inefficiency and congestion with associated measures in DEA, *Socio-Economic Planning Sciences*, 34, pp. 1–25.

Cooper, W.W., Thompson, R.G. & Thrall, R.M. (1996) Introduction: Extensions and new developments in DEA, *Annals of Operations Research*, 66, pp. 3–45.

Färe, R. & Grosskopf, S. (2000a) Slacks and congestion: a comment, *Socio-Economic Planning Sciences*, 34, pp. 27–33.

Färe, R. & Grosskopf, S. (2000b) Research note: Decomposing technical efficiency with care, *Management Science*, 46, pp. 167–168.

Färe, R., Grosskopf, S., Lindgren, B. & Roos, P. (1992) Productivity changes in Swedish pharmacies 1980–1989: a non-parametric Malmquist approach, *Journal of Productivity Analysis*, 3, pp. 85–101.

Färe, R., Grosskopf, S. & Logan, J. (1985a) The relative performance of publicly-owned and privately-owned electric utilities, *Journal of Public Economics*, 26, pp. 89–106.

Färe, R., Grosskopf, S. & Lovell, C.A.K. (1985b) *The Measurement of Efficiency of Production* (Boston, Kluwer-Nijhoff)

Färe, R., Grosskopf, S., Norris, M. & Zhang, Z. (1994) Productivity growth, technical progress, and efficiency change in industrialized countries, *American Economic Review*, 84, pp. 66–83.

Farrell, M.J. (1957) The measurement of productive efficiency, *Journal of the Royal Statistical Society*, Series A, General, 120, Part 3, pp. 253–281.

Flegg, A.T., Allen, D.O., Field, K. & Thurlow, T.W. (2004) Measuring the Efficiency of British Universities: A Multi-Period Data Envelopment Analysis, *Education Economics*, 12, pp. 231–249.

Tone, K. & Sahoo, B.K. (2004) Degree of scale economies and congestion: a unified DEA approach, *European Journal of Operational Research*, 158, pp. 755–772.

University Statistics (various years) *University Statistics*, various issues (Cheltenham, Universities' Statistical Record).

Table 1. Comparing different means for each approach

Färe's approach: C_F						Ranking of means			
	AM	**WAM**	**GM**	**WGM**	**sd**	**AM**	**WAM**	**GM**	**WGM**
1980	1.0657	1.0655	1.0617	1.0614	0.002	1	2	3	4
1981	1.0598	1.0602	1.0561	1.0565	0.002	2	1	4	3
1982	1.0635	1.0715	1.0599	1.0670	0.005	3	1	4	2
1983	1.0594	1.0631	1.0547	1.0580	0.003	2	1	4	3
1984	1.0932	1.0973	1.0811	1.0835	0.008	2	1	4	3
1985	1.0671	1.0813	1.0579	1.0700	0.010	3	1	4	2
1986	1.0641	1.0719	1.0576	1.0644	0.006	3	1	4	2
1987	1.0518	1.0612	1.0485	1.0574	0.006	3	1	4	2
1988	1.0419	1.0435	1.0391	1.0400	0.002	2	1	4	3
1989	1.0611	1.0682	1.0559	1.0622	0.005	3	1	4	2
1990	1.0384	1.0430	1.0366	1.0410	0.003	3	1	4	2
1991	1.0526	1.0535	1.0504	1.0513	0.001	2	1	4	3
1992	1.0326	1.0356	1.0316	1.0344	0.002	3	1	4	2
sd	0.0153	0.0166	0.0127	0.0134					
Tone's approach: C_T						Ranking of means			
	AM	**WAM**	**GM**	**WGM**	**sd**	**AM**	**WAM**	**GM**	**WGM**
1980	1.0611	1.0625	1.0571	1.0584	0.002	2	1	4	3
1981	1.0545	1.0562	1.0509	1.0526	0.002	2	1	4	3
1982	1.0570	1.0661	1.0537	1.0619	0.005	3	1	4	2
1983	1.0520	1.0554	1.0475	1.0505	0.003	2	1	4	3
1984	1.0880	1.0925	1.0764	1.0793	0.008	2	1	4	3

1985	1.0661	1.0801	1.0570	1.0689	0.010	3	1	4	2
1986	1.0618	1.0696	1.0553	1.0620	0.006	3	1	4	2
1987	1.0489	1.0591	1.0458	1.0555	0.006	3	1	4	2
1988	1.0364	1.0385	1.0337	1.0351	0.002	2	1	4	3
1989	1.0557	1.0642	1.0507	1.0583	0.006	3	1	4	2
1990	1.0337	1.0393	1.0321	1.0373	0.003	3	1	4	2
1991	1.0492	1.0509	1.0471	1.0487	0.002	2	1	4	3
1992	1.0286	1.0321	1.0275	1.0309	0.002	3	1	4	2
sd	0.0154	0.0168	0.0127	0.0137					
Cooper's approach: $1 + C_C$						Ranking of means			
	AM	**WAM**	**GM**	**WGM**	**sd**	**AM**	**WAM**	**GM**	**WGM**
1980	1.0518	1.0527	1.0502	1.0506	0.001	2	1	4	3
1981	1.0362	1.0419	1.0352	1.0403	0.003	3	1	4	2
1982	1.0384	1.0398	1.0373	1.0385	0.001	3	1	4	2
1983	1.0340	1.0397	1.0331	1.0386	0.003	3	1	4	2
1984	1.0636	1.0720	1.0602	1.0679	0.005	3	1	4	2
1985	1.0459	1.0552	1.0435	1.0520	0.005	3	1	4	2
1986	1.0423	1.0538	1.0403	1.0511	0.007	3	1	4	2
1987	1.0412	1.0549	1.0390	1.0517	0.008	3	1	4	2
1988	1.0494	1.0489	1.0471	1.0463	0.001	1	2	3	4
1989	1.0471	1.0610	1.0450	1.0579	0.008	3	1	4	2
1990	1.0350	1.0393	1.0335	1.0375	0.003	3	1	4	2
1991	1.0480	1.0624	1.0459	1.0592	0.008	3	1	4	2
1992	1.0447	1.0557	1.0431	1.0531	0.006	3	1	4	2
sd	0.0081	0.0100	0.0075	0.0092					

Table 2. Ranking the results from alternative approaches

		Approach			**Ranking of approach**				**Ranking of year**		
		Färe (F)	**Tone (T)**	**Cooper (C)**	**F**	**T**	**C**	**range**	**F**	**T**	**C**
Arithmetic Mean (AM)	1980	1.0657	1.0611	1.0518	1	2	3	0.0139	3	4	2
	1981	1.0598	1.0545	1.0362	1	2	3	0.0236	7	7	11
	1982	1.0635	1.0570	1.0384	1	2	3	0.0251	5	5	10
	1983	1.0594	1.0520	1.0340	1	2	3	0.0254	8	8	**13**
	1984	1.0932	1.0880	1.0636	1	2	3	0.0296	**1**	**1**	**1**
	1985	1.0671	1.0661	1.0459	1	2	3	0.0212	2	2	6
	1986	1.0641	1.0618	1.0423	1	2	3	0.0218	4	3	8
	1987	1.0518	1.0489	1.0412	1	2	3	0.0106	10	10	9
	1988	1.0419	1.0364	1.0494	2	3	1	0.0129	11	11	3
	1989	1.0611	1.0557	1.0471	1	2	3	0.0140	6	6	5
	1990	1.0384	1.0337	1.0350	1	3	2	0.0046	12	12	12
	1991	1.0526	1.0492	1.0480	1	2	3	0.0046	9	9	4
	1992	1.0326	1.0286	1.0447	2	3	1	0.0161	**13**	**13**	7
W	1980	1.0655	1.0625	1.0527	1	2	3	0.0128	6	6	8
	1981	1.0602	1.0562	1.0419	1	2	3	0.0183	9	8	10
	1982	1.0715	1.0661	1.0398	1	2	3	0.0317	4	4	11
	1983	1.0631	1.0554	1.0397	1	2	3	0.0233	7	9	12
	1984	1.0973	1.0925	1.0720	1	2	3	0.0253	**1**	**1**	**1**
	1985	1.0813	1.0801	1.0552	1	2	3	0.0261	2	2	5
	1986	1.0719	1.0696	1.0538	1	2	3	0.0181	3	3	7
	1987	1.0612	1.0591	1.0549	1	2	3	0.0063	8	7	6
	1988	1.0435	1.0385	1.0489	2	3	1	0.0104	11	12	9
	1989	1.0682	1.0642	1.0610	1	2	3	0.0072	5	5	3
	1990	1.0430	1.0393	1.0393	1	2	3	0.0038	12	11	**13**
	1991	1.0535	1.0509	1.0624	2	3	1	0.0115	10	10	2

	1992	1.0356	1.0321	1.0557	2	3	1	0.0236	**13**	**13**	4
Geometric Mean (GM)	1980	1.0617	1.0571	1.0502	1	2	3	0.0115	2	2	2
	1981	1.0561	1.0509	1.0352	1	2	3	0.0209	6	6	11
	1982	1.0599	1.0537	1.0373	1	2	3	0.0226	3	5	10
	1983	1.0547	1.0475	1.0331	1	2	3	0.0216	8	8	**13**
	1984	1.0811	1.0764	1.0602	1	2	3	0.0209	**1**	**1**	**1**
	1985	1.0579	1.0570	1.0435	1	2	3	0.0144	4	3	6
	1986	1.0576	1.0553	1.0403	1	2	3	0.0173	5	4	8
	1987	1.0485	1.0458	1.0390	1	2	3	0.0095	10	10	9
	1988	1.0391	1.0337	1.0471	2	3	1	0.0134	11	11	3
	1989	1.0559	1.0507	1.0450	1	2	3	0.0110	7	7	5
	1990	1.0366	1.0321	1.0335	1	3	2	0.0045	12	12	12
	1991	1.0504	1.0471	1.0459	1	2	3	0.0045	9	9	4
	1992	1.0316	1.0275	1.0431	2	3	1	0.0156	**13**	**13**	7
Weighted Geometric Mean (WGM)	1980	1.0614	1.0584	1.0506	1	2	3	0.0108	6	5	8
	1981	1.0565	1.0526	1.0403	1	2	3	0.0163	9	8	10
	1982	1.0670	1.0619	1.0385	1	2	3	0.0285	3	4	12
	1983	1.0580	1.0505	1.0386	1	2	3	0.0195	7	9	11
	1984	1.0835	1.0793	1.0679	1	2	3	0.0156	**1**	**1**	**1**
	1985	1.0700	1.0689	1.0520	1	2	3	0.0180	2	2	5
	1986	1.0644	1.0620	1.0511	1	2	3	0.0133	4	3	7
	1987	1.0574	1.0555	1.0517	1	2	3	0.0057	8	7	6
	1988	1.0400	1.0351	1.0463	2	3	1	0.0112	12	12	9
	1989	1.0622	1.0583	1.0579	1	2	3	0.0043	5	6	3
	1990	1.0410	1.0373	1.0375	1	3	2	0.0037	11	11	**13**
	1991	1.0513	1.0487	1.0592	2	3	1	0.0104	10	10	2
	1992	1.0344	1.0309	1.0531	2	3	1	0.0222	**13**	**13**	4

Table 3. Scale diseconomies: Tone's approach

	$\bar{\rho}$ for all universities n = 45	Year rank n = 45	Number of congested universities	Year rank	Congested universities			
					$\bar{\rho}$	Max	Min	V
1980	–1.662	7	24	6	–3.117	–13.29	–0.132	3.183
1981	–2.455	10	26	9	–4.249	–19.39	–0.033	4.139
1982	–3.139	**13**	24	**13**	–5.885	–20.01	–0.334	5.341
1983	–3.003	12	25	12	–5.405	–28.24	–0.156	7.307
1984	–2.520	11	24	11	–4.725	–31.67	–0.182	6.023
1985	–1.356	5	21	5	–2.905	–6.68	–0.660	1.751
1986	–1.350	4	21	4	–2.893	–6.41	–0.867	1.571
1987	–1.473	6	21	7	–3.157	–7.41	–0.732	1.721
1988	–1.929	8	23	8	–3.773	–25.37	–0.899	5.153
1989	–0.829	**1**	22	**1**	–1.695	–3.76	–0.135	0.860
1990	–1.124	3	19	3	–2.663	–4.29	–0.986	1.057
1991	–2.455	9	24	10	–4.602	–38.44	–0.903	7.683
1992	–0.961	2	25	2	–1.730	–4.73	–0.163	1.250

Note: V is the coefficient of variation.

Table 4a. Sources of congestion: Cooper's approach

$C_C = [S_1/X_1 + S_2/X_2 + S_3/X_3 + S_4/X_4]/4$ (annual average per input, n = 45)						
	Undgrad S_1/X_1	Postgrad S_2/X_2	Ac staff S_3/X_3	Dep exp S_4/X_4	C_C	rank
1980	0.1028	0.0295	0.0644	0.0106	0.0518	2
1981	0.0667	0.0223	0.0443	0.0113	0.0362	11
1982	0.0736	0.0205	0.0449	0.0145	0.0384	10
1983	0.0715	0.0163	0.0366	0.0115	0.0340	**13**
1984	0.1241	0.0423	0.0835	0.0046	0.0636	**1**
1985	0.0882	0.0368	0.0587	0.0000	0.0459	6
1986	0.0949	0.0239	0.0473	0.0029	0.0423	8
1987	0.0736	0.0270	0.0566	0.0076	0.0412	9
1988	0.0768	0.0385	0.0755	0.0068	0.0494	3
1989	0.0929	0.0243	0.0684	0.0028	0.0471	5
1990	0.0683	0.0363	0.0334	0.0021	0.0350	12
1991	0.0993	0.0463	0.0423	0.0042	0.0480	4
1992	0.0752	0.0546	0.0462	0.0027	0.0447	7

Table 4b. Decomposition of overall congestion score

Contribution of inputs (%)				
	Undgrad S_1/X_1	Postgrad S_2/X_2	Ac staff S_3/X_3	Dep exp S_4/X_4
1980	49.57	14.23	31.08	5.12
1981	46.12	15.42	30.64	7.82
1982	47.95	13.33	29.27	9.44
1983	52.65	11.99	26.92	8.45
1984	48.76	16.62	32.82	1.80
1985	48.03	20.04	31.94	0.00

1986	56.14	14.14	27.99	1.72
1987	44.67	16.37	34.37	4.59
1988	38.89	19.48	38.21	3.42
1989	49.31	12.90	36.28	1.51
1990	48.74	25.91	23.84	1.50
1991	51.70	24.10	22.02	2.18
1992	42.07	30.54	25.86	1.53
Min	38.89	11.99	22.02	0.00
Max	56.14	30.54	38.21	9.44
Mean	48.05	18.08	30.10	3.78

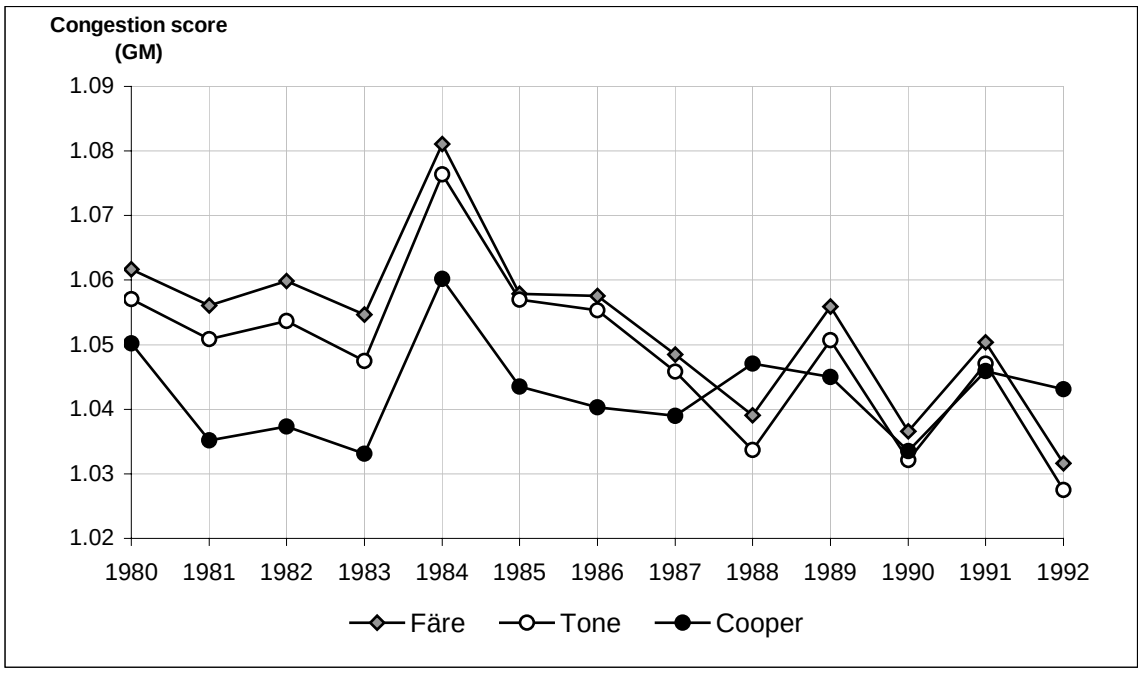

Fig 6a. Comparing congestion scores: Geometric Mean (GM).

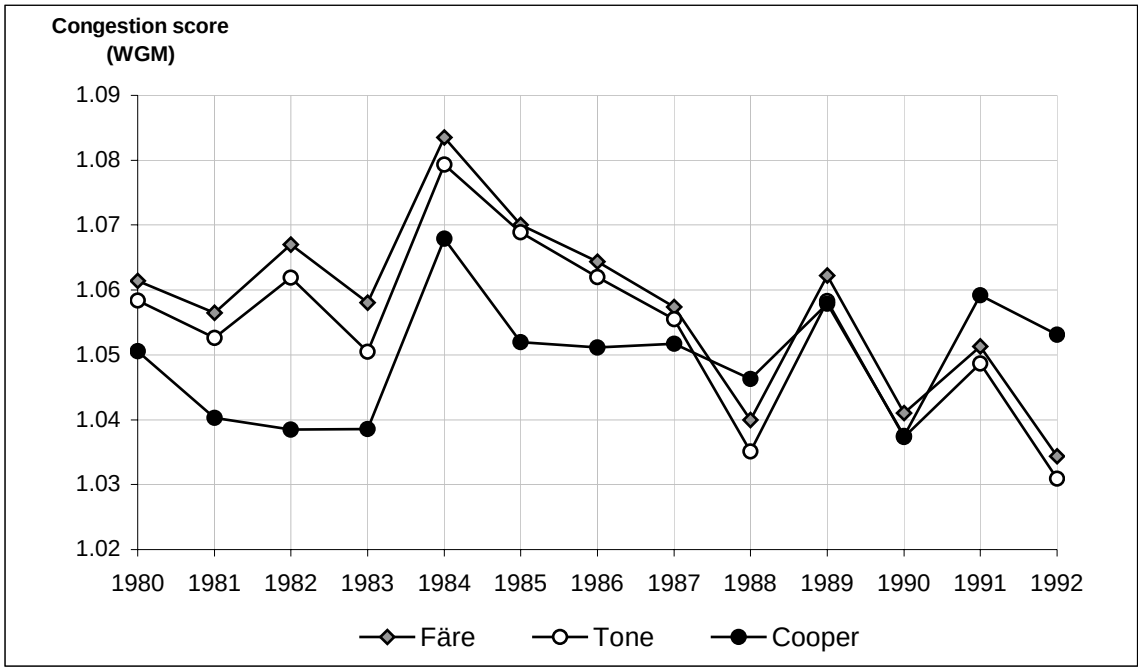

Fig. 6b. Comparing congestion scores: Weighted Geometric Mean (WGM).

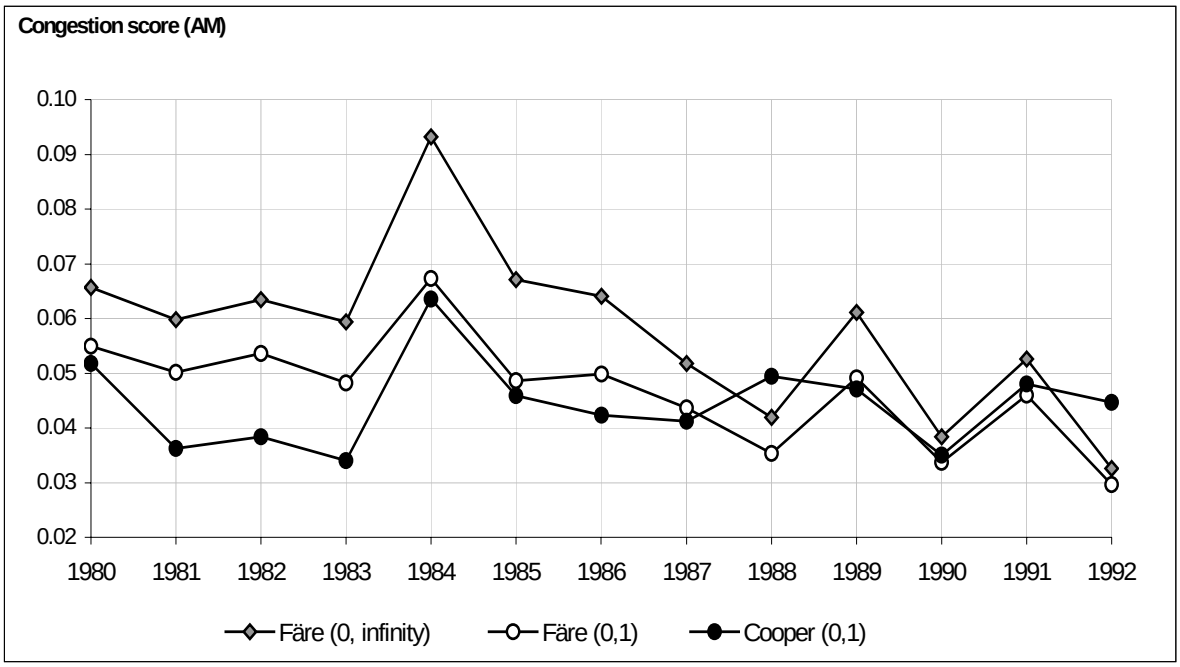

Fig. 7a. **Comparing congestion scores: Arithmetic Mean (AM).**

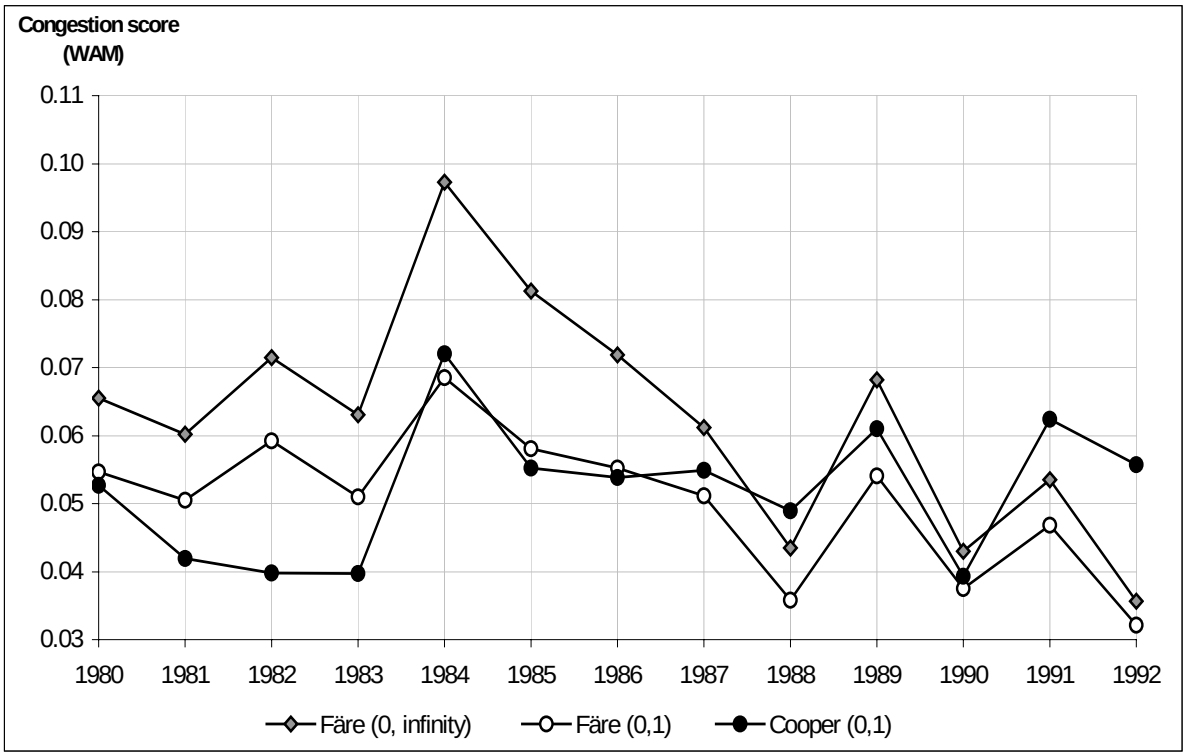

Fig. 7b. Comparing congestion scores: Weighted Arithmetic Mean (WAM).

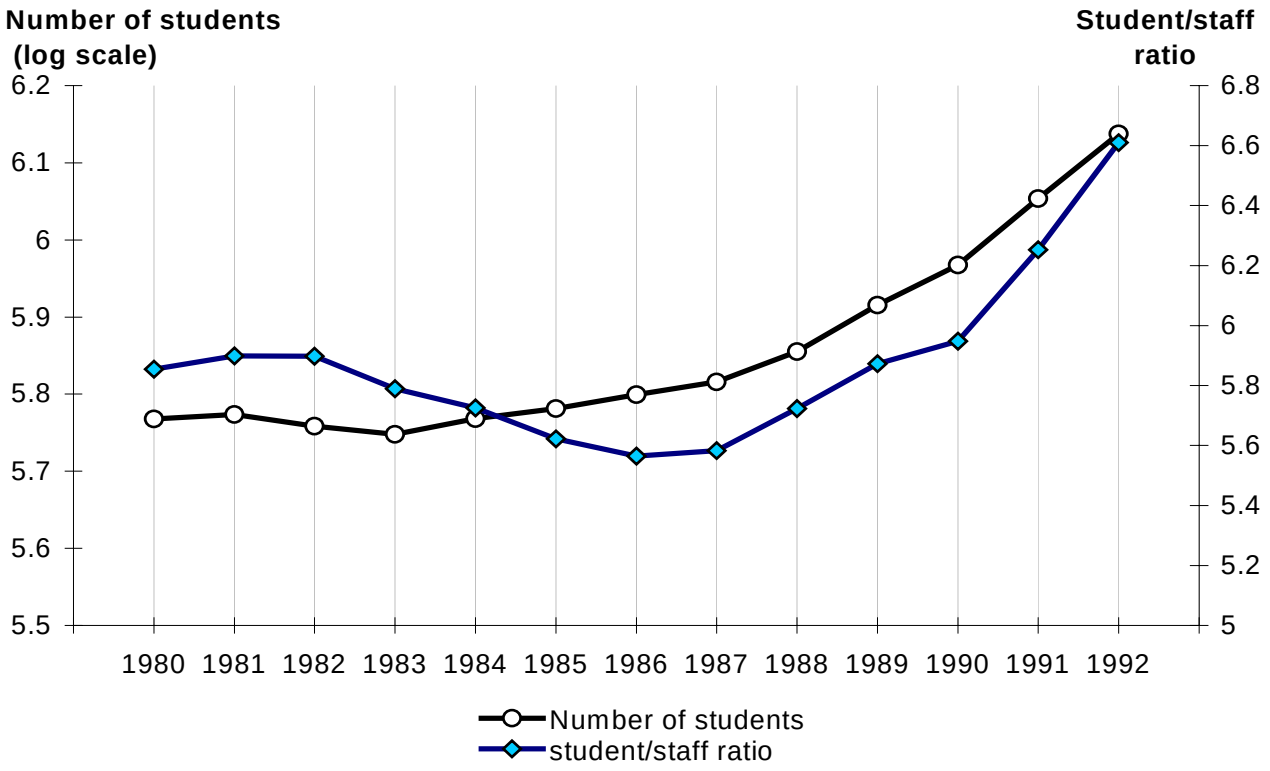

Fig. 1. Students and staff: UK universities, 1980/81–1992/93.

Notes: Derived from data in *University Statistics* (various years). See Flegg *et al.* (2004, p. 2

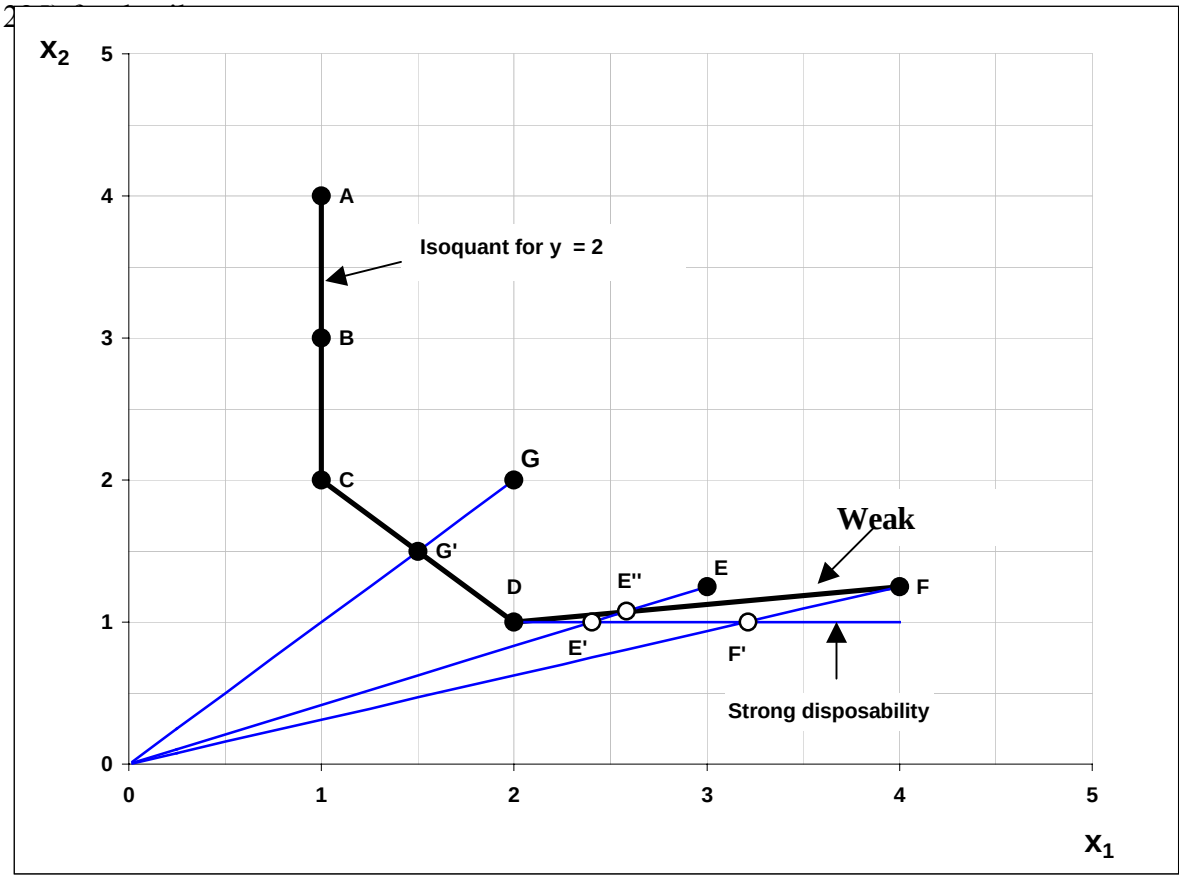

Fig. 2. Färe's approach (input-oriented, CRS).

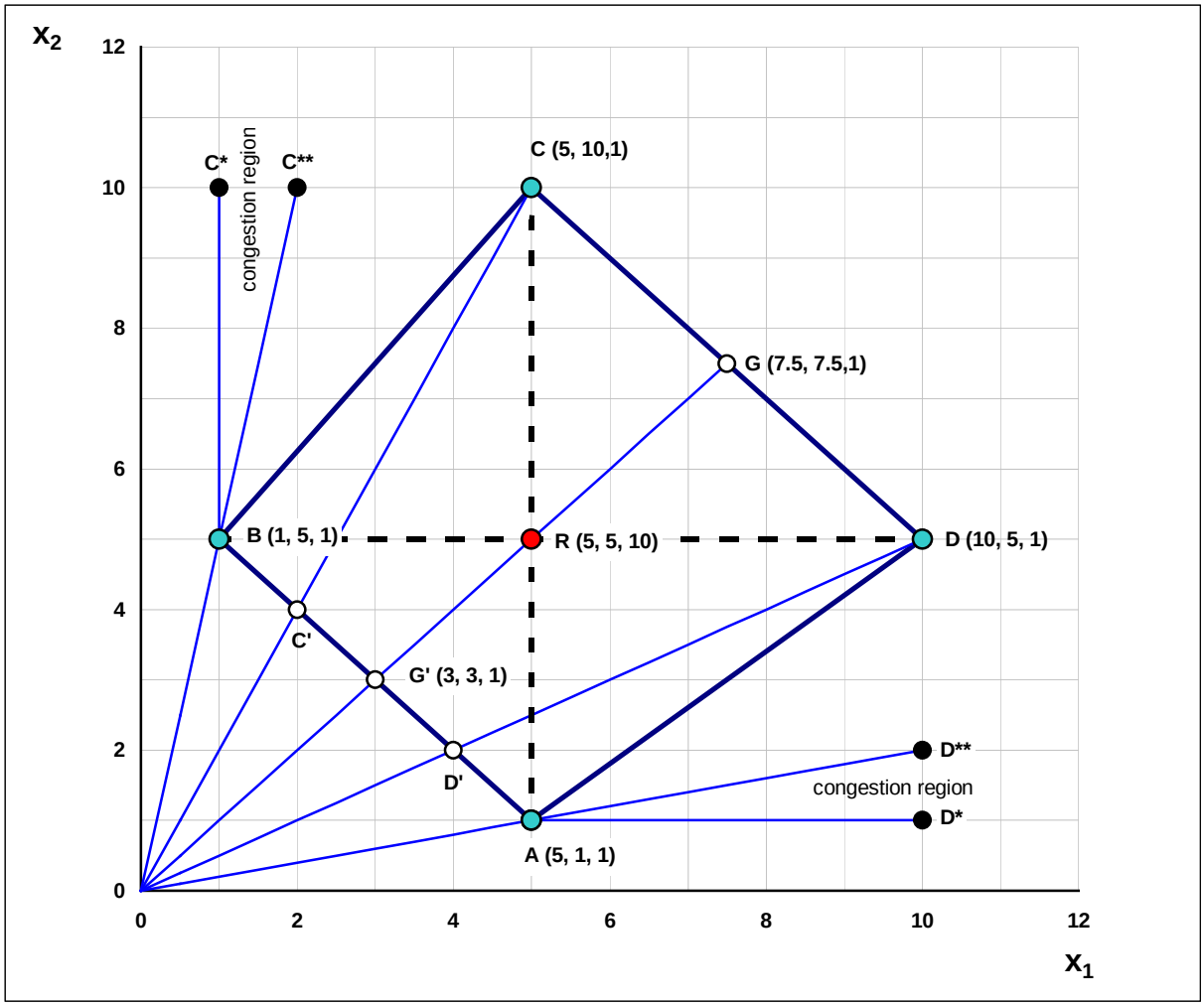

Fig. 3. Färe's approach (input-oriented, VRS).

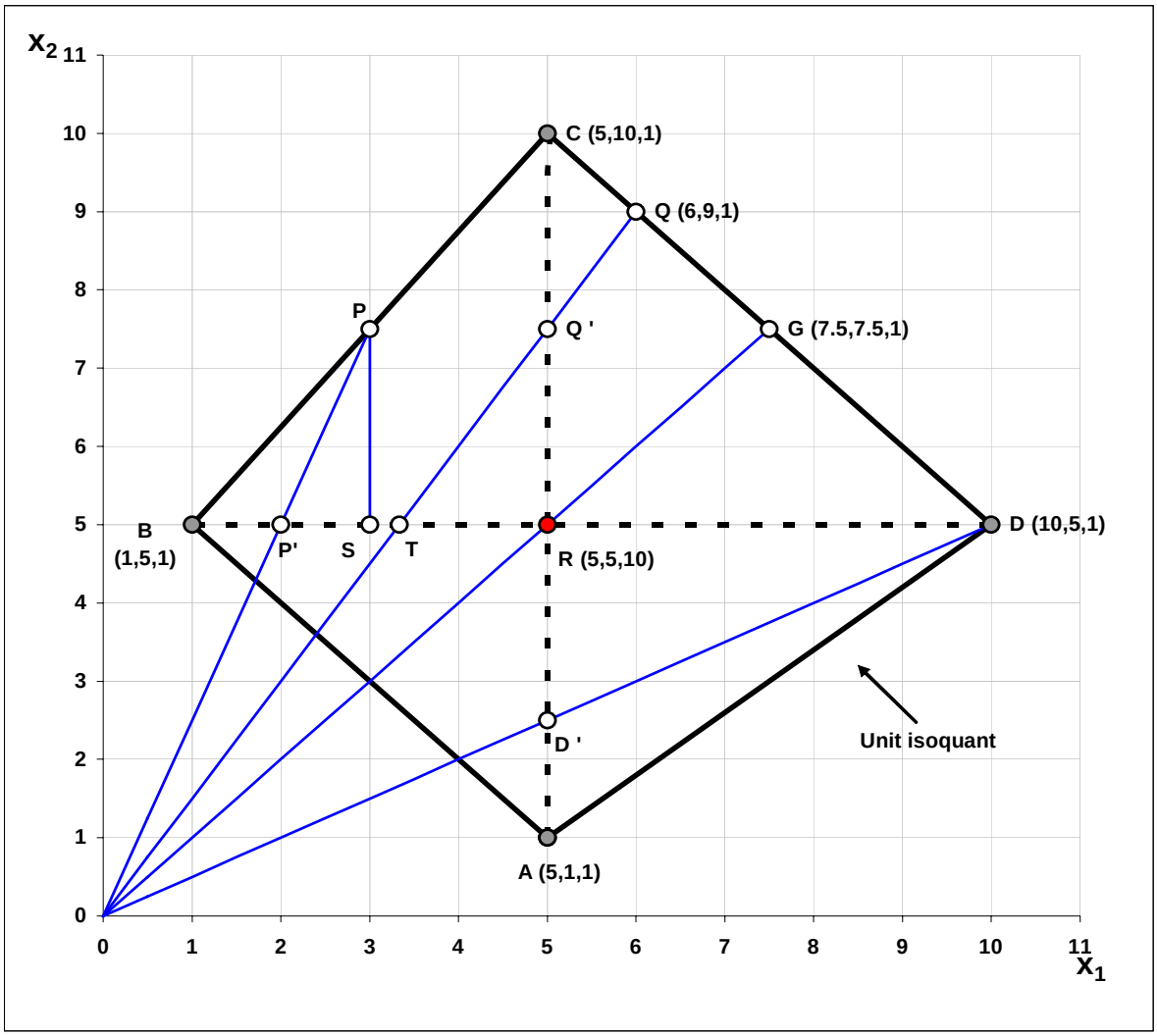

Fig. 4. Färe's approach (output-oriented, VRS).

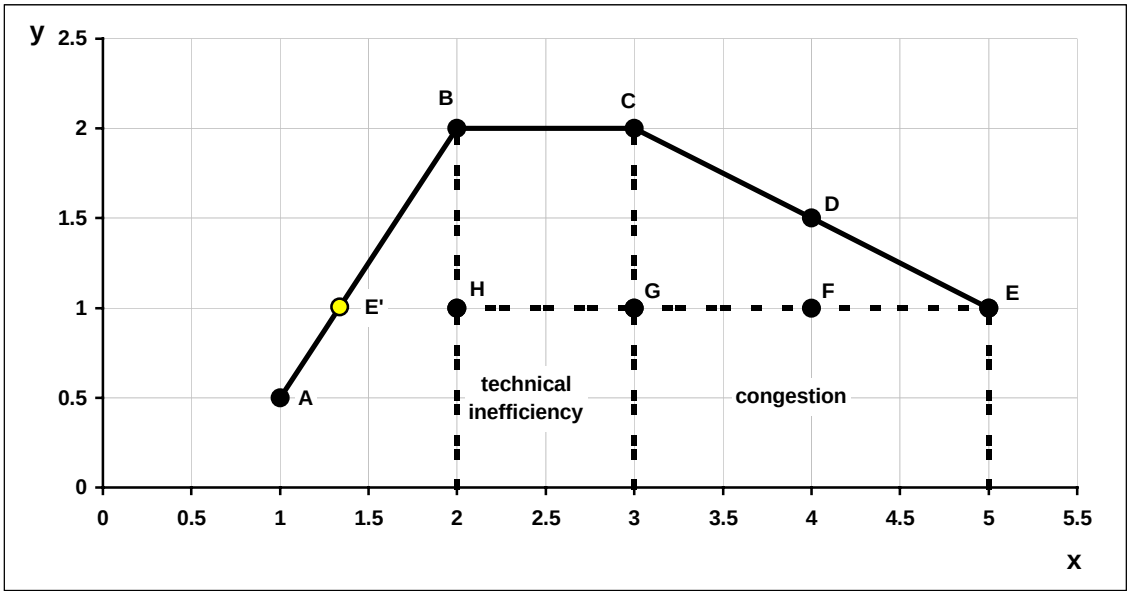

Fig. 5. Cooper's output-oriented approach.