An Examination of Alternative Approaches to Measuring Congestion in British Universities

A.T. FLEGG & D.O. ALLEN*

ABSTRACT This paper examines three alternative methods of measuring congestion, from both theoretical and empirical perspectives. These methods are 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 to 1992/93. Despite conceptual differences, Tone and Sahoo's approach and that of Cooper *et al*. are found to produce fairly similar results. Contrary to expectations, Färe and Grosskopf's approach generally indicates more congestion than the other two procedures. The main reason for this is identified as being its use of CRS rather than VRS as the assumed technology. Although the three alternative measures of congestion are found to be positively correlated, the correlations are not strong enough for them to be regarded as substitutes. Also contrary to expectations, the results suggest that academic overstaffing, rather than excessive numbers of undergraduates, was the largest single cause of congestion in British universities during the period under review. Even so, only a modest amount of congestion is identified.

Key words: British universities Congestion DEA

In a recent paper in *Education Economics* (Flegg *et al.*, 2004), we examine the impact on British universities' efficiency of the rapid and unbalanced expansion in the period 1980/81 to 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 the 45 universities suffered from congestion in the sense that they could have produced a larger output

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

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 to 1992/93. In this regard, it is worth noting Färe and Grosskopf's observation that, of the two procedures, their own approach would generally indicate *less* congestion.

In this re-examination of our earlier study, we employ an input-oriented rather than output-oriented version of Färe and Grosskopf's procedure. We also incorporate some refinements of the variables used in our earlier study and carry out a sensitivity analysis. In addition, we 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. What is Congestion?

Cooper, Gu and Li (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, Figure 1 shows that there was a seemingly inexorable rise in the student : staff ratio in British universities from 1986/87 onwards; 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.

2. Measuring Congestion

The conventional way of measuring congestion was developed by Färe and Grosskopf, while Byrnes, Färe and Grosskopf (1984) and Färe, Grosskopf and Logan (1985) were the first published applications. Cooper, Thompson and Thrall (1996) then proposed an alternative procedure, which was refined and applied to Chinese data by Brockett *et al.* (1998) and by Cooper, Seiford and Zhu (2000). The merits and demerits of the two approaches have been debated most recently by Cherchye, Kuosmanen and Post (2001) and Cooper, Gu and Li (2001a, 2001b). 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, Grosskopf and Lovell, 1985). 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. As illustrated later, an input exhibits 'slack' in situations where it is possible to reduce the quantity used of that input without causing output to decline. Färe ignores such slacks on the basis that they can be disposed of at no 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, Gu and Li (2001a, p. 69) posit the following relationship (the notation has been simplified):

(1) $\quad c_i = s_i^{-*} - \delta_i^*$

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, Gu and Li 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 illustrated best by the use of examples. However, before examining these examples, we should note that is possible to decompose Färe and Grosskopf's measure of overall technical efficiency (TE) in a straightforward way into pure technical efficiency (PTE), scale efficiency (SE) and congestion efficiency (CE), using the identity:

(2) $\quad TE \equiv PTE \times SE \times CE$

where TE = 1 and TE < 1 represent technical efficiency and inefficiency, respectively.

2.1. Example 1 (see Figure 2)

Figure 2 shows six decision-making units (DMUs), each producing an output of y = 1. This example assumes *constant returns to scale* (CRS), so that SE = 1, and makes use of an *input-oriented* approach.[1] As regards D and E, there would be no dispute between the two schools of thought: both DMUs are clearly technically efficient. Likewise, there would be agreement concerning C's technical inefficiency. In terms of identity (2), TE = PTE = ⅔ for C. However, under Färe's approach, F would also be deemed to be efficient. Färe would

[1] A similar example is discussed in Cooper, Gu and Li (2001a). The views attributed here to Cooper mirror those of Cooper, Gu and Li.

disregard the fact that F has slack in x_1 of 1 unit. By contrast, Cooper would treat this DMU as being only *weakly* efficient. Whereas Cooper regards slack as a form of technical inefficiency, Färe argues that 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 A and B. Because A is on the isoquant for y = 1, Färe would regard this DMU as exhibiting no *pure* technical inefficiency (PTE = 1). However, it does appear to suffer from congestion. Its CE score, as measured by the ratio OA′/OA, equals 0.8. Its 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 DA, which is assumed to exhibit *weak* disposability, and the hypothetical vertical line emanating from D, which is assumed to exhibit *strong* (or free) disposability. By moving to point A′, A could attain TE = 1. This would be the end of the matter according to Färe. However, Cooper would then point to the slack in x_2 of DA′ = 1.2 units and say that this was indicative of technical inefficiency but not congestion.

The case of B is more complicated because Färe would claim that this DMU suffers from both pure technical inefficiency and congestion. PTE and CE are measured, respectively, by the ratios OB′′/OB ≈ 0.714 and OB′/OB′′ ≈ 0.933. Hence TE = ⅔ ≈ 0.714 × 0.933. Färe would ignore the slack in x_2 of DB′ = ⅓ of a unit.

By contrast, Cooper would assert that there is no evidence that either A or B suffers from congestion! This is because all DMUs in Figure 2 produce the same output of y = 1. 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 C to B, raising the quantity of x_2 by 0.5, 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 slack of A and B to technical inefficiency (δ_i^* in equation (1) above): $\delta_2^* = 1.2$ for A and $0.\dot{3}$ for B.

In the context of this example, however, these criticisms of Färe's approach 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 C from 1 to, say, 1.25 but leaving the output of all other DMUs constant at 1. If we now moved from C to B, the rise in x_2 from 3 to 3.5 would be accompanied by a *fall* in output from 1.25 to 1. Clearly, this would constitute 'congestion' in the sense of Definition 1 above.

What is more, even if all DMUs had y = 1, we could still validly argue that A and B suffered from congestion in input x_2. This is because, along segment DA, the marginal product of x_2 must be negative. Output stays constant along DA because the rise due to greater use of the non-congested input x_1 exactly offsets the fall due to greater use of the congested input x_2.

2.2. 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, Gu and Li (2001a). Whereas R produces an output of y = 10, the remaining DMUs all produce y = 1. The figure takes the form of a pyramid with its pinnacle at R. *Variable returns to scale* (VRS) are now assumed. Given this assumption, A and B are efficient.[2] However, under Färe's approach, 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 PTE = 0.4 and CE = 1.

Cooper would dispute the finding of no congestion in the case of G; indeed, he would argue that there is, in fact, compelling evidence of its existence. For instance, suppose that

[2] A and B would be inefficient (TE < 1) under CRS whereas R would be efficient.

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, Gu and Li (2001a, 2001b) 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, Gu and Li (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, Kuosmanen and Post (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 D and C, in turn.

If we move D to position D* in Figure 3, then PTE = CE = 1. There is thus no congestion. This is also true if we move D to position D**, although 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 C* and C** or in between D* and D** has $0.5 < CE < 1$.

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 apparent problems with Färe's approach, we might ask whether Cooper's approach would fare any better. Cooper, Gu and Li (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.

Since the above discussion has revealed some potential problems with using input-oriented models, it is logical to consider the use of an output-oriented approach. Unfortunately, as the examples under consideration here involve a single output, it would not be appropriate to

consider an output-oriented version of Färe's approach.[3] Therefore, in the next example, we will confine ourselves to examining an output-oriented version of Cooper's approach.

2.3. Example 3 (see Figure 4)

Before examining Figure 4, 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:

(3) $\quad c_i/x_i = s_i^{-*}/x_i - \delta_i^*/x_i$

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:[4]

(4) $\quad C_C = \overline{s/x} - \overline{\delta/x}$

Hence C_C measures the average proportion of congestion in the inputs used by a particular DMU. It has the property $0 \leq C_C \leq 1$. See Cooper, Gu and Li (2001a, p. 73).

The first stage of Cooper's procedure makes use of the *output-oriented* version of the Banker–Charnes–Cooper (BCC) model. This, in turn, involves two steps. In the first step, the model below is employed to obtain the value of ϕ^* for each DMU k, while the second step involves maximizing the sum of the slacks, conditional on this value of ϕ^* (cf. Cooper, Seiford and Zhu, 2000, pp. 3–5):

(5a) $\quad \phi^* = \max \phi$

subject to:

(5b) $\quad \sum_j \lambda_j x_{ij} \leq x_{ik} \qquad i = 1, 2, \ldots, m$

(5c) $\quad \sum_j \lambda_j y_{rj} \geq \phi y_{rk} \qquad r = 1, 2, \ldots, s$

[3] We are indebted to Pontus Roos of the Economic Measurement and Quality (EMQ) Corporation in Sweden for pointing this out.

[4] There is a case for using geometric rather than arithmetic means to average these ratios.

(5d) $\sum_j \lambda_j = 1$

(5e) $\lambda_j \geq 0$ $\quad j = 1, 2, \ldots, n$

To illustrate the use of Cooper's model, consider DMU E in Figure 4. This diagram reveals that there are two possible referent DMUs available for evaluating 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 4, we would move along the BCC frontier from B to C, holding output constant at y = 2. This process would yield $\delta_x^* = 1$.

Thus, in the case of 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 (4), 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 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 E, thereby again illustrating the disadvantages of this orientation when measuring congestion of inputs (the projection would have been to point E′ in Figure 4).

In real data sets, horizontal segments such as BC in Figure 4 are rare and, in our own data set of 45 universities over 13 years, no case occurs where $\phi^* = 1$, yet non-zero slack exists. 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 3 to see how Cooper's approach would evaluate the DMUs shown there. In the case of G, we get $C_C = ½\{(2.5/7.5) + (2.5/7.5)\} = ⅓$. $C_C = 0.25$ for C and D. There is, therefore, a modest rise in the measured amount of congestion as we approach G from either side. As regards segment BC, the value of C_C rises monotonically from zero at B to reach a maximum of 0.25 at C. The same thing happens along segment AD.

If we accept – as the present authors do – that all points (apart from A and B) lying on the segments BC, CD and AD of the frontier in Figure 3 are congested (since the marginal product of x_1 or x_2 or both is negative), then the output-oriented version of Cooper's procedure is clearly able to identify the congestion that exists.[5] In contrast, with Färe's approach, a DMU located at any one of these points 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. However, we would readily acknowledge the hazards of generalizing from a particular numerical example about the relative performance of different approaches (cf. Cherchye, Kuosmanen and Post, 2001, p. 76). There are also some other considerations, discussed below, that need to be borne in mind when choosing a particular method of identifying and measuring congestion.

3. 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 in a straightforward way into pure technical efficiency, scale efficiency and congestion efficiency, using the identity (2) above. 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. This software also makes it possible

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

to select – on *a priori* grounds – which inputs are to be examined for possible congestion. 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, Kuosmanen and Post (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 Figure 3 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.

The most attractive feature of Cooper's approach is that it makes use of concepts that can easily 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, Seiford and Zhu, 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 (4) 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 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, Seiford and Zhu (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.[6] In the present paper, we aim to add to the scant empirical evidence on this topic.

4. 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.[7] 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.

[6] It is worth noting that, when computing Färe's measures, Cooper *et al.* assumed VRS rather than CRS. Their study also involved a single output and time-series data, whereby each year was treated as a separate DMU. By contrast, our own study employs panel data and several outputs.

[7] We are indebted to Kaoru Tone for confirming this point.

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 3.

Like Cooper, Tone would find A, B and R to be BCC efficient and hence not congested. The remaining DMUs would have a congestion score of $\theta = 10$, 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:

$$(6) \qquad \rho = \frac{\%\text{ change in y}}{\%\text{ change in x}_2} = \frac{+900\%}{-50\%} = -18$$

Using the same method, we also get $\rho = -18$ for D. In the case of G, the average percentage change in inputs is $-33\frac{1}{3}\%$, so that $\rho = -27$. 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 $C_C = \frac{1}{3}$ 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).

5. The Model and Methodology

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

- income from research grants and contracts and from other services rendered;
- the number of undergraduate degrees awarded, adjusted for quality;[8]
- the number of postgraduate degrees awarded.

[8] 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 full-time equivalent undergraduate students (X_1);
- the number of full-time equivalent postgraduate students (X_2);
- the number of academic and academic-related staff (X_3);[9]
- aggregate departmental recurrent expenditure (X_4).[10]

This model is referred to here as Model 1. A rationale for the variables is given in the paper cited above. Data were obtained from *University Statistics* (various years).

In our earlier study, an *output-oriented* variant of Färe's approach was used to compute a congestion efficiency score for each university. A weighted mean was then calculated for each year, using the number of students in each university as a weight, to take account of the diverse size of universities. Here we have modified our use of Färe's approach to take account of recent theoretical developments.

The first issue concerns the *order* in which technical efficiency (TE) is decomposed into pure technical efficiency (PTE), scale efficiency (SE) and congestion efficiency (CE). In their earlier work, Färe and Grosskopf assumed strong disposability when measuring scale effects, and only then allowed for the possibility of congestion.[11] However, Färe and Grosskopf (2000b) have highlighted the problems associated with distinguishing between scale inefficiency and congestion; they point out that the CE score will depend on the order in which TE is decomposed.[12] Therefore, where congestion is anticipated on *a priori* grounds, Färe and Grosskopf recommend that one should base one's measurements on CRS rather than on VRS technology. We have followed this suggestion here.

[9] Part-time staff were given a weight of 0.5.

[10] This includes expenditure on: salaries and wages of *non-academic* staff, equipment, research grants and contracts, along with some other unspecified items.

[11] See, for example, Byrnes, Färe and Grosskopf (1984), and Färe, Grosskopf and Logan (1985).

[12] In the identity TE ≡ PTE × SE × CE, TE and the product SE × CE are unaffected by the order of the decomposition but the individual values of SE and CE are affected.

The other issue concerns the *orientation* of the model and the distinction between input and output congestion. In the current version of *OnFront*, congestion of inputs is measured using an *input-oriented* approach, whereas congestion of outputs is captured via an *output-oriented* approach. In the case of outputs, congestion refers to a situation where one or more of the outputs is an undesirable by-product of joint production, e.g. air pollution associated with the generation of electricity. Since all three outputs in our model are deemed to be desirable, congestion of outputs can be ruled out *a priori*. On the other hand, there are sound reasons for anticipating congestion with respect to one or more of the inputs.

In view of the above arguments, we will be employing an *input-oriented* variant of Färe's approach, with CRS as the underlying technology, to compute a CE score for each university. This approach is consistent with the earlier discussion surrounding Example 1 and Figure 2. However, we will revisit this issue of the underlying technology later in the paper.

6. Initial Results

The top panel of Table 1 shows the annual unweighted arithmetic mean (UAM) congestion scores for the three approaches: F for Färe, T for Tone and C for Cooper. The bottom panel shows the corresponding weighted arithmetic mean (WAM) scores. The number of students in each university was used as a weight. These results are illustrated in Figure 5.

For Cooper's approach, the mean scores were calculated by first working out C_C, the average proportion of congestion in the inputs used by each university in each year, and then averaging these figures over the 45 universities. For consistency with Cooper's measure, the congestion efficiency (CE) scores from Färe's input-oriented approach were converted into *in*efficiency scores, viz $C_F \equiv 1 - CE$, before averaging over all universities. In the case of Tone's output-oriented approach, the following transformation was used: $C_T \equiv 1 - 1/\theta$, where

$\theta \geq 1$ is the congestion score generated by *DEA-Solver Pro*.[13] With these transformations, all measures have a convenient range from 0 (no congestion) to 1 (maximum congestion).

A striking facet of the unweighted results is the fact that Färe's measure, $\overline{C}_F$, clearly signals the highest amount of congestion in the first eight years, yet his measure is the only one indicating a clear tendency for congestion to diminish over the period. By contrast, Cooper's measure, $\overline{C}_C$, ends up a little higher in 1992/93 than in 1980/81, whereas the opposite is true of Tone's measure, $\overline{C}_T$. It is worth noting that $\overline{C}_F$ invariably exceeds $\overline{C}_T$, although these two measures tend to converge towards the end of the period. Indeed, all three measures are much closer at the end of the period than at the start. Even so, Cooper's measure still ranks 1980/81 as the least congested year, whereas the other two measures opt for 1992/93.

However, when the congestion scores are weighted by the number of students in each university, a rather different picture emerges. As can be seen from Table 1 and Figure 5, the three measures are now in agreement that congestion fell over the period, albeit by a modest amount. It is interesting that the three measures now start at almost the same point in 1980/81; what is more, they end up in 1992/93 in the same order and with almost exactly the same values as they did in the unweighted analysis. Nonetheless, one can see that the use of weights causes Cooper's measure to be even more volatile from 1988/89 onwards. In general, the weighting has the effect of lowering the values of $\overline{C}_F$ but raising the values of $\overline{C}_C$ and $\overline{C}_T$.[14]

The correlation coefficients presented in Table 2 shed some further light on the relationships among the three measures. These correlations are based on the raw scores, so that $n = 13 \times 45 = 585$. Looking at the results for Model 1, one can see that the three

[13] An alternative would be to define Tone's measure as $C_T \equiv \theta - 1$. Cooper *et al.* (2000) followed this approach when transforming Färe's *output-oriented* measure to enable comparisons to be made with C_C. However, measures of this kind have no finite upper limit and their use could distort comparisons with measures constrained to a [0, 1] range. A demerit of using a [0, 1] range is that geometric means cannot be used, as they were in our earlier study, when averaging the congestion scores.

[14] A higher WAM than UAM suggests that relatively large universities were more congested than relatively small ones.

measures are positively correlated. C_T and C_C are the most strongly correlated, whereas C_C and C_F have the weakest correlation. However, in no case is the correlation strong enough for the measures to be regarded as substitutes. As expected, all measures are negatively correlated with TE; this finding suggests that a reduction in congestion is associated with enhanced technical efficiency. However, none of these correlations is particularly strong.

It is also informative to see how far the measures agree on which universities are congested and which are not. In fact, of the 585 cases examined, $C_C = C_T = 1$ in 293. Only two instances were discovered of $C_C < 1$ with $C_T = 1$ and none of $C_C = 1$ with $C_T < 1$. These two measures thus show a remarkable degree of accord in classifying universities. Even so, the correlation coefficient of 0.731 between C_C and C_T suggests some degree of disagreement over the severity of this congestion in particular universities. It is also possible that there is an element of non-linearity in the relationship between C_C and C_T. As regards Färe's measure, this suggests a more widespread problem of congestion: $C_F = 1$ in only 204 cases. Even so, in all but nine of these cases, it was also true that $C_F = C_T = 1$, which indicates some degree of similarity between the two measures.

How can we explain this close matching of the universities deemed to be congested by Cooper and Tone? An important point to note here is that both approaches use an output-oriented version of the BCC model as their starting point. Thus scale effects are removed prior to attempting to measure congestion. Also, only those universities deemed to be inefficient in terms of the BCC model are examined for possible congestion. Therefore, even though Cooper and Tone measure congestion somewhat differently, they are still looking at the same set of universities. By contrast, Färe's measure employs an input-oriented approach with CRS as the underlying technology. In addition, it uses 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. Cooper's method is also non-radial. Hence it is not surprising that the results from the different methods do not coincide.

7. A New Model

The model examined thus far (Model 1) has the number of first-class honours degrees as the measure of undergraduate output. We adopted this measure in our earlier study in an attempt to control for the *quality* of undergraduate degrees. However, a shortcoming of this measure is its narrowness. A better measure would have comprised the number of firsts and upper seconds but this information was not published in *University Statistics* for the period under review. A second problem is that it is, perhaps, rather anomalous to adjust the output of undergraduate degrees for quality but not the input of undergraduates! For these reasons, we have developed a new model (Model 2), in which undergraduate output is measured by the number of undergraduate degrees awarded, irrespective of classification. We have also added post-graduate certificates and diplomas (such as post-graduate certificates of education) to our measure of post-graduate output.

Before looking at the congestion scores generated by Model 2, it may be helpful to examine the impact on technical efficiency of using the broader measures of undergraduate and postgraduate output. Figure 6 shows the weighted mean TE scores for each model.[15] It is immediately apparent that the mean scores are substantially higher for Model 2. This is not surprising inasmuch as the use of broader measures of output should make it easier for universities to improve their relative performance. Consequently, we are likely to observe less *variation* in performance in each year and hence obtain higher values for $\overline{\text{TE}}$. It is also evident that there is a clear upward trend in the values of $\overline{\text{TE}}$ for Model 1 but no trend in the case of Model 2. As a result, the gap between the two graphs decreases noticeably over the period.

The use of Model 2 also results in a substantial fall in the weighted arithmetic mean congestion scores from all three approaches. This can be seen by comparing the corresponding

[15] The graph shown in Figure 6 for Model 1 is somewhat smoother than the corresponding WAM graph in Flegg *et al.* (2004, Figure 1). This is because we were able to remove two anomalies in the data: an inconsistency over the whole period in the allocation of postgraduate degrees to academic years and an apparent understatement of the number of postgraduate degrees awarded by Hull in 1983/84.

results in Tables 1 and 3, which are illustrated in Figure 7. This decrease in the values of $\overline{C}_F$, $\overline{C}_C$ and $\overline{C}_T$ is not surprising, given the substantial decline in the values for $\overline{TE}$. It is also evident that there is now much less variation in the values of each measure over the period. This, again, is to be expected.

An interesting finding from Model 2 is that Färe's measure clearly suggests that the problem of congestion is more widespread than is indicated by the other two measures. This is consistent with the outcome from Model 1, where an analysis of the raw scores revealed that Färe's measure generated a noticeably larger number of congested universities. Another noteworthy finding is that the three measures are in accord that congestion *rose* over the period as a whole, albeit not by very much. This conclusion is the opposite of that reached in the case of Model 1!

The correlations shown in Table 2 for Models 1 and 2 are broadly similar. As before, there is a negative correlation of each measure with TE; in addition, the three measures are positively correlated, yet not strongly enough for them to be treated as substitutes. The correlation between C_T and C_C is now noticeably weaker than it was for Model 1, although C_C and C_F again have the weakest correlation.

An analysis of the raw scores for Model 2 revealed that $C_T = 1$ in 397 (67.9%) of the 585 cases considered, while $C_C = 1$ in 384 (65.6%). What is more, $C_T = C_C = 1$ in all of these 384 cases. The thirteen exceptions had $C_T = 1$ but $C_C < 1$. Färe's measure produced only 264 cases (45.1%) of no congestion, of which 260 had $C_F = C_T = 1$. By comparison, for Model 1, far fewer cases were identified of no congestion: 293 (50.1%) for Cooper, 295 (50.4%) for Tone and 204 (34.9%) for Färe.

8. Scale Diseconomies

In addition to generating congestion scores, Tone's approach also provides some useful information on scale diseconomies. Table 4 shows the annual arithmetic mean values of ρ,

Tone's *scale diseconomies* parameter, based on Model 1 and 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 9.2% on average in 1984/85 but only 1.9% in 1989/90. This suggests that congestion was much more serious in 1984/85. 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 $\overline{\rho}$ suggests that congestion was most serious in 1984/85 but least serious in 1989/90, $\overline{C}_T$ picks out 1981/82 as the year with the most congestion and 1992/93 as the year with the least. At first sight, this disagreement is somewhat surprising. However, the differences in the values of $\overline{\rho}$ for 1989/90 and 1992/93 are rather small. As regards 1984/85, this year produced an exceptionally large value of $\rho = -129.3$ for Aberdeen; excluding this university had the effect of lowering $|\rho|$ from 9.16 to 3.69. The value of $\overline{\rho}$ for 1983/84 was also unduly influenced by another unusually large value of $\rho = -73.3$ for Aberdeen.

An examination of the output from *DEA-Solver Pro* revealed that ρ was much more prone than C_T to fluctuate from year to year.[16] For example, for Aberdeen, $|\rho|$ rose dramatically from zero in 1982/83 to 129.3 in 1984/85, whereas C_T rose more gently from zero to 0.114. A less dramatic example is Reading, where $|\rho|$ fell sharply from 27.7 in 1982/83 to 1.63 in 1984/85, whereas C_T fell less noticeably from 0.066 to 0.043. In both cases, the congestion was associated with a large shortfall in the number of first-class degrees awarded.

Table 5 shows the corresponding results from Model 2. As expected, the values of $\overline{\rho}$ are much smaller than those for Model 1. In terms of scale diseconomies, congestion was most

[16] ρ has a much larger coefficient of variation (V) than C_T. For n = 45, the value of V for C_T ranged from 0.048 to 0.154 over the study period. Unlike C_T, ρ has no upper bound, and hence is likely to be more volatile as a result.

serious in 1982/83 but least serious in 1986/87. $\overline{C}_T$ also ranks 1986/87 as the least congested year, although it picks out 1981/82 as the most congested year.

It is clear that $\overline{C}_T$ and $\overline{\rho}$ are unlikely to yield the same ranking of years in terms of congestion. Given its sensitivity to extreme values, $\overline{\rho}$ is not a very reliable measure of the amount of congestion in a given year. Nonetheless, ρ does provide some very useful information about potential scale diseconomies in individual universities.

9. Sources of Congestion

A helpful aspect of Cooper's approach is that it permits one to examine, for each university, the contribution of each input to the observed amount of congestion. Tables 6 to 9 take a closer look at this facet of Cooper's method. Tables 6 and 8 show, for each model, how $\overline{C}_C$ was calculated in each year (using unweighted arithmetic means), while Tables 7 and 9 show the contribution of each input to the value of $\overline{C}_C$.

The results for Model 1 (Table 7) reveal that excessive numbers of undergraduates (X_1) were the largest single cause of congestion in British universities, accounting for between 32% and 55% 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 (X_3) accounted for a higher proportion of $\overline{C}_C$ than did undergraduates. The table indicates that postgraduates (X_2) typically had a substantially smaller role than undergraduates in terms of causing congestion. Finally, we can see that excessive departmental expenditure (X_4) was of only minor importance.

Strikingly different results are obtained when Model 2 replaces Model 1. Table 9 shows a huge reduction in the proportion of congestion that can be attributed to excessive numbers of undergraduates. There is also a big rise in the amount of congestion due to academic

overstaffing (especially from 1986/87) and to excessive departmental expenditure. Postgraduates are relatively unaffected by the change in model.

It is easy to explain the smaller role of undergraduates in generating congestion. In Model 1, the only undergraduate output recognized is the award of a first-class degree, whereas all undergraduate degrees are recognized as being equally valid in Model 2. This change obviously reduces the scope for observing 'excessive' numbers of undergraduates! With regard to the role of postgraduates, little change is observed because the impact of including certificates and diplomas in the postgraduate output variable was relatively small.

The finding regarding academic overstaffing is puzzling – especially in view of the sharp rise in the student : staff ratio from 1986/87 onwards (see Figure 1)! What this finding 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, as well as undergraduate and postgraduate qualifications 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.

10. Order of Decomposition

The results from Model 2 revealed some similarity between the approaches of Tone and Cooper, with Färe's approach standing out as being the most different. Indeed, Table 3 and Figure 7 show that, of the three measures, C_F is the one indicating the most congestion.

Hitherto, Färe's measure of congestion, C_F, has been calculated by using CRS as the underlying technology. This is the approach recommended by Färe and Grosskopf (2000b) in cases where congestion is anticipated on *a priori* grounds. By contrast, Cooper and Tone use VRS as the underlying technology when measuring congestion. Therefore, to explore this issue, we recalculated C_F using VRS.[17] The unweighted results for Model 2 are shown in Figure 8.

A glance at Figure 8 is all that is needed to see that we get far less 'congestion' if we assume VRS rather than CRS. What is more, Figure 9 shows clearly that there is now little difference between Färe's measure and that of Tone. Even though the deviations are fairly small, it is now Cooper's approach that stands out as the most different from the other two.

The correlations shown in Table 10 substantiate the point that, if we assume VRS, Färe's measure is strongly correlated with that of Tone.[18] The fact that this correlation is 0.886 rather than unity can be attributed to two factors: the use of an input orientation and the different ways in which congestion is measured. As regards the orientation, we were rather concerned about employing an input orientation when using Färe's measure. This is because we would 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. It is reassuring, therefore, that the orientation does not seem to be a major factor in explaining the divergence between the measures of Färe and Tone.[19]

It is also worth noting that, if we assume VRS, there is a very close correspondence between the sets of universities deemed to be congested under each approach. An analysis of

[17] In cases where congestion is anticipated, Färe and Grosskopf (2000b) recommend that one should compare a (CRS, S) model with a (CRS, W) model, as opposed to comparing a (VRS, S) model with a (VRS, W) model (where S = strong disposability and W = weak disposability).

[18] C_F is also more strongly correlated with C_C than it was previously (r = 0.619 versus 0.487).

[19] In our earlier study, Flegg *et al.* (2004), we used an output-oriented variant of Färe's approach and assumed VRS. When we reworked the results using Tone's approach, we got remarkably similar congestion scores.

the raw scores for Model 2 revealed that $C_T = 1$ in 397 (67.9%) of the 585 cases considered, whereas $C_F = 1$ in 388 (66.3%). There were only 19 cases of disagreement as to whether a particular university was or was not congested in a given year.

Conclusion

This paper has examined three alternative methods of measuring congestion, from both theoretical and empirical perspectives. These methods were the well-known procedure of Färe and Grosskopf, the alternative approach developed by Cooper and his associates, and a new method proposed by Tone and Sahoo.

The theoretical discussion identified some apparent 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 score of a given university. What is more, this score can easily be decomposed into scale, congestion and purely 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. The approach is also well supported by the *OnFront* software.

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 shed light on the underlying causes of congestion. 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 with 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 methods were compared using annual data for 45 British universities over the period 1980/81 to 1992/93. An examination of the raw data for individual universities revealed that the measures of congestion were positively correlated but that the correlations were not strong enough for them to be regarded as substitutes. However, Tone and Sahoo's method and that of Cooper *et al.* did produce remarkably similar results in terms of identifying which universities were congested and which were not, although they differed with regard to the severity of the congestion in each congested university. It is worth noting that Färe and Grosskopf's method generated a substantially larger number of congested universities than either of the other two procedures.

The analysis was conducted in terms of two alternative models, which differed in terms of how broadly or narrowly the measures of undergraduate and postgraduate output were defined. In Model 1, the only undergraduate output recognized was the award of a first-class degree, whereas all undergraduate degrees were recognized as being equally valid in Model 2. With regard to postgraduates, the models differed in that postgraduate certificates and diplomas were included in the postgraduate output variable employed in Model 2.

Strikingly different results were obtained from the two models. The weighted mean technical efficiency score was substantially higher for Model 2 than for Model 1, and the weighted mean congestion scores were substantially lower. What is more, whereas Model 1 produced a substantial rise in technical efficiency and a concomitant fall in congestion over the period 1980/81 to 1992/93, Model 2 indicated no trend in technical efficiency and a slight rise in congestion!

Contrary to expectations, the results for Model 2 revealed that academic overstaffing, rather than excessive numbers of undergraduates, was the largest single cause of congestion in British universities during the period 1980/81 to 1992/93. On average, academic staff accounted for 43% of the value of Cooper's congestion score. By contrast, excessive

departmental expenditure accounted for 24%. The remaining 33% was attributed, in almost equal measure, to postgraduates and undergraduates.

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. It has to be said, however, that the overall amount of congestion indicated by Model 2 was modest, ranging over the study period from 0.8% of inputs, on average, to 1.7%. By contrast, the figures for Model 1 ranged from 2.9% to 6.4%.

The results revealed that the three alternative measures of congestion are not close substitutes, so that it would be wise to consider their properties carefully before opting for any one of them. Indeed, since each measure has its merits and demerits, it seems sensible not to rely on the use of a single procedure. A crucial issue appears to be the *order* in which technical efficiency is decomposed into scale and congestion components. Here Färe and Grosskopf recommend using CRS as the underlying technology in cases where congestion is anticipated on *a priori* grounds. Based on our findings, if one followed their recommendation, one could expect to uncover more 'congestion' than if one used VRS. If Färe and Grosskopf are correct, then some of the scale inefficiency indicated by the other two methods should, in fact, be regarded as congestion. Another issue concerns the orientation of the model. Here the results suggested that this issue may be less crucial in practice than whether one opts for CRS or VRS.

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. Alternative measures of congestion: Model 1

	Unweighted Arithmetic Mean (UAM)				Ranking by method			Ranking of year		
	Färe	**Tone**	**Cooper**	**Mean**	**F**	**T**	**C**	**F**	**T**	**C**
1980	0.062	0.036	0.029	0.042	1	2	3	8	5	**1**
1981	0.064	0.059	0.040	0.054	1	2	3	10	**13**	6
1982	0.073	0.044	0.036	0.051	1	2	3	**13**	8	3
1983	0.060	0.039	0.039	0.046	1	2	3	7	6	4
1984	0.064	0.054	0.056	0.058	1	3	2	9	12	12
1985	0.065	0.051	0.040	0.053	1	2	3	11	11	5
1986	0.069	0.046	0.041	0.053	1	2	3	12	9	8
1987	0.057	0.036	0.042	0.045	1	3	2	5	4	9
1988	0.059	0.046	0.064	0.056	2	3	1	6	10	**13**
1989	0.044	0.041	0.041	0.042	1	2	3	4	7	7
1990	0.035	0.031	0.044	0.037	2	3	1	3	2	10
1991	0.035	0.032	0.045	0.037	2	3	1	2	3	11
1992	0.032	0.027	0.036	0.031	2	3	1	**1**	**1**	2
Min	0.032	0.027	0.029	0.031						
Max	0.073	0.059	0.064	0.058						
Mean	0.055	0.042	0.043	0.047						
SD	0.014	0.010	0.009	0.008						

	Weighted Arithmetic Mean (WAM)				Ranking by method			Ranking of year		
	Färe	**Tone**	**Cooper**	**Mean**	**F**	**T**	**C**	**F**	**T**	**C**
1980	0.043	0.041	0.040	0.041	1	2	3	5	5	3
1981	0.048	0.060	0.044	0.051	2	1	3	7	12	6
1982	0.058	0.048	0.040	0.049	1	2	3	10	9	2
1983	0.047	0.043	0.042	0.044	1	2	3	6	6	4
1984	0.058	0.056	0.060	0.058	2	3	1	11	11	12
1985	0.061	0.063	0.055	0.060	2	1	3	12	**13**	9
1986	0.066	0.050	0.052	0.056	1	3	2	**13**	10	8
1987	0.053	0.039	0.045	0.046	1	3	2	8	4	7
1988	0.054	0.048	0.072	0.058	2	3	1	9	8	**13**
1989	0.041	0.047	0.044	0.044	3	1	2	4	7	5
1990	0.031	0.037	0.060	0.043	3	2	1	**1**	3	11
1991	0.036	0.033	0.059	0.042	2	3	1	3	2	10
1992	0.032	0.025	0.036	0.031	2	3	1	2	**1**	**1**
Min	0.031	0.025	0.036	0.031						
Max	0.066	0.063	0.072	0.060						
Mean	0.048	0.045	0.050	0.048						
SD	0.011	0.011	0.011	0.008						

Table 2. Correlation of measures of congestion
(unweighted, n = 585)

Model 1			
	TE	C_F	C_T
C_F	–0.598		
C_T	–0.613	0.673	
C_C	–0.546	0.512	0.731
Model 2			
	TE	C_F	C_T
C_F	–0.604		
C_T	–0.528	0.630	
C_C	–0.450	0.487	0.584

Table 3. Alternative measures of congestion: Model 2

	Unweighted Arithmetic Mean (UAM)				Ranking by method			Ranking of year		
	Färe	**Tone**	**Cooper**	**Mean**	**F**	**T**	**C**	**F**	**T**	**C**
1980	0.020	0.009	0.008	0.012	1	2	3	7	6	3
1981	0.024	0.016	0.012	0.017	1	2	3	10	**13**	8
1982	0.025	0.009	0.012	0.015	1	3	2	12	5	7
1983	0.030	0.014	0.015	0.020	1	3	2	**13**	11	11
1984	0.025	0.015	0.015	0.019	1	3	2	11	12	12
1985	0.021	0.004	0.010	0.012	1	3	2	9	2	5
1986	0.018	0.003	0.004	0.008	1	3	2	3	**1**	**1**
1987	0.020	0.010	0.012	0.014	1	3	2	6	9	9
1988	0.021	0.009	0.011	0.014	1	3	2	8	8	6
1989	0.017	0.009	0.014	0.013	1	3	2	2	4	10
1990	0.014	0.006	0.009	0.010	1	3	2	**1**	3	4
1991	0.019	0.009	0.008	0.012	1	2	3	4	7	2
1992	0.020	0.013	0.017	0.017	1	3	2	5	10	**13**
Min	0.014	0.003	0.004	0.008						
Max	0.030	0.016	0.017	0.020						
Mean	0.021	0.010	0.011	0.014						
SD	0.004	0.004	0.004	0.003						

	Weighted Arithmetic Mean (WAM)				Ranking by method			Ranking of year		
	Färe	**Tone**	**Cooper**	**Mean**	**F**	**T**	**C**	**F**	**T**	**C**
1980	0.014	0.005	0.007	0.009	1	3	2	2	3	3
1981	0.018	0.012	0.010	0.013	1	2	3	9	11	6
1982	0.021	0.009	0.010	0.013	1	3	2	10	9	7
1983	0.029	0.014	0.015	0.019	1	3	2	**13**	**13**	11
1984	0.024	0.012	0.015	0.017	1	3	2	12	12	12
1985	0.022	0.005	0.011	0.012	1	3	2	11	2	8
1986	0.016	0.002	0.003	0.007	1	3	2	4	**1**	**1**
1987	0.018	0.007	0.010	0.012	1	3	2	8	7	5
1988	0.016	0.008	0.011	0.012	1	3	2	6	8	9
1989	0.015	0.007	0.015	0.012	1	3	2	3	6	**13**
1990	0.012	0.005	0.007	0.008	1	3	2	**1**	4	4
1991	0.016	0.006	0.005	0.009	1	2	3	5	5	2
1992	0.017	0.010	0.014	0.013	1	3	2	7	10	10
Min	0.012	0.002	0.003	0.007						
Max	0.029	0.014	0.015	0.019						
Mean	0.018	0.008	0.010	0.012						
SD	0.004	0.003	0.004	0.003						

Table 4. Scale diseconomies: Model 1

	All universities (n = 45)				Congested universities				
	$\overline{C}_T$	Rank	$\bar{\rho}$	Rank	Number	$\bar{\rho}$	Max	Min	V
1980	0.036	5	–1.68	9	20	–3.78	–10.3	–1.24	2.61
1981	0.059	**13**	–3.04	11	28	–4.89	–19.4	–0.60	4.08
1982	0.044	8	–2.68	10	22	–5.48	–27.7	–0.18	5.83
1983	0.039	6	–3.28	12	22	–6.72	–73.3	–0.66	15.49
1984	0.054	12	–4.68	**13**	23	–9.16	–129.3	–0.51	26.36
1985	0.051	11	–1.42	6	20	–3.20	–5.7	–1.01	1.46
1986	0.046	9	–1.14	3	18	–2.85	–5.0	–1.08	1.14
1987	0.036	4	–1.43	7	23	–2.79	–7.0	–0.28	1.71
1988	0.046	10	–1.43	8	25	–2.58	–7.3	–0.50	1.81
1989	0.041	7	–0.88	**1**	21	–1.88	–4.5	–0.08	1.10
1990	0.031	2	–1.25	4	22	–2.56	–6.8	–0.67	1.35
1991	0.032	3	–1.40	5	24	–2.63	–8.7	–0.62	1.97
1992	0.027	**1**	–1.02	2	22	–2.08	–5.5	–0.29	1.55

Table 5. Scale diseconomies: Model 2

	All universities (n = 45)				Congested universities				
	$\overline{C}_T$	Rank	$\bar{\rho}$	Rank	Number	$\bar{\rho}$	Max	Min	V
1980	0.009	6	–1.17	12	11	–4.77	–37.6	–0.07	11.00
1981	0.016	**13**	–0.72	5	12	–2.68	–11.3	–0.49	2.92
1982	0.009	5	–1.80	**13**	17	–4.77	–50.3	–0.30	11.78
1983	0.014	11	–0.82	7	15	–2.47	–7.6	–0.15	2.29
1984	0.015	12	–0.84	10	18	–2.09	–6.0	–0.03	1.72
1985	0.004	2	–1.15	11	12	–4.32	–14.5	–0.36	5.25
1986	0.003	**1**	–0.28	**1**	8	–1.57	–5.0	–0.60	1.44
1987	0.010	9	–0.83	8	19	–1.96	–9.7	–0.24	2.54
1988	0.009	8	–0.83	9	16	–2.34	–6.5	–0.61	1.44
1989	0.009	4	–0.68	4	18	–1.69	–6.2	–0.28	1.48
1990	0.006	3	–0.34	2	10	–1.52	–4.0	–0.41	1.10
1991	0.009	7	–0.78	6	12	–2.93	–22.6	–0.17	6.26
1992	0.013	10	–0.59	3	20	–1.32	–4.1	–0.01	1.13

Table 6. Components of Cooper's congestion score (Model 1, n = 45)

$\overline{C}_C = [S_1/X_1 + S_2/X_2 + S_3/X_3 + S_4/X_4]/4$							
	Ugrads S_1/X_1	Pgrads S_2/X_2	Staff S_3/X_3	Expend S_4/X_4	$\overline{C}_C$	Rank	Number congested
1980	0.055	0.011	0.046	0.004	0.029	**1**	20
1981	0.068	0.031	0.046	0.017	0.040	6	28
1982	0.075	0.010	0.046	0.013	0.036	3	22
1983	0.085	0.018	0.042	0.009	0.039	4	23
1984	0.106	0.038	0.077	0.002	0.056	12	23
1985	0.080	0.029	0.045	0.005	0.040	5	21
1986	0.089	0.019	0.054	0.003	0.041	8	18
1987	0.080	0.015	0.065	0.008	0.042	9	23
1988	0.082	0.072	0.095	0.007	0.064	**13**	25
1989	0.078	0.021	0.062	0.003	0.041	7	21
1990	0.082	0.039	0.052	0.002	0.044	10	22
1991	0.091	0.039	0.045	0.005	0.045	11	24
1992	0.054	0.040	0.043	0.007	0.036	2	22

Table 7. Percentage contribution of each input: Model 1

	Ugrads X_1	Pgrads X_2	Staff X_3	Expend X_4
1980	47.2	9.1	39.9	3.8
1981	42.3	19.1	28.2	10.4
1982	51.7	7.1	31.9	9.3
1983	55.3	11.8	27.2	5.7
1984	47.6	17.0	34.3	1.1
1985	50.5	18.2	28.3	2.9
1986	54.0	11.6	32.4	2.0
1987	47.8	8.8	38.8	4.6
1988	32.1	28.2	37.0	2.7
1989	47.6	12.6	38.2	1.6
1990	46.8	22.1	29.8	1.2
1991	50.5	21.8	25.2	2.5
1992	37.6	27.5	29.9	5.0
Min	32.1	7.1	25.2	1.1
Max	55.3	28.2	39.9	10.4
Mean	47.0	16.5	32.4	4.1
SD	6.4	7.0	4.9	3.0

Table 8. Components of Cooper's congestion score (Model 2, n = 45)

$\overline{C}_C = [S_1/X_1 + S_2/X_2 + S_3/X_3 + S_4/X_4]/4$							
	Ugrads S_1/X_1	Pgrads S_2/X_2	Staff S_3/X_3	Expend S_4/X_4	$\overline{C}_C$	Rank	Number congested
1980	0.001	0.003	0.016	0.012	0.008	3	12
1981	0.007	0.003	0.018	0.020	0.012	8	14
1982	0.007	0.002	0.018	0.020	0.012	7	18
1983	0.015	0.009	0.020	0.017	0.015	11	17
1984	0.014	0.008	0.021	0.019	0.015	12	18
1985	0.010	0.005	0.010	0.015	0.010	5	12
1986	0.002	0.002	0.010	0.004	0.004	**1**	8
1987	0.010	0.007	0.028	0.005	0.012	9	20
1988	0.001	0.009	0.025	0.011	0.011	6	17
1989	0.007	0.005	0.037	0.008	0.014	10	18
1990	0.007	0.011	0.017	0.000	0.009	4	14
1991	0.010	0.010	0.008	0.002	0.008	2	13
1992	0.016	0.021	0.024	0.007	0.017	**13**	20

Table 9. Percentage contribution of each input: Model 2

	Ugrads X_1	Pgrads X_2	Staff X_3	Exp X_4
1980	4.0	9.9	49.0	37.0
1981	14.0	6.8	36.8	42.5
1982	14.6	5.2	38.4	41.9
1983	24.5	14.9	32.5	28.1
1984	22.6	13.3	33.4	30.8
1985	24.5	12.9	24.6	38.0
1986	8.8	11.9	54.9	24.3
1987	20.3	13.1	55.8	10.8
1988	1.6	19.0	55.2	24.2
1989	12.7	8.6	64.1	14.6
1990	19.3	31.0	49.7	0.0
1991	32.1	33.0	26.8	8.0
1992	24.0	30.6	34.7	10.7
Min	1.6	5.2	24.6	0.0
Max	32.1	33.0	64.9	42.5
Mean	17.2	16.2	42.8	23.9
SD	8.9	9.5	12.6	14.0

Table 10. Correlation of measures of congestion
(Model 2, unweighted, n = 585)

	TE	$C_{F, CRS}$	$C_{F, VRS}$	C_T
$C_{F, CRS}$	−0.604			
$C_{F, VRS}$	−0.567	0.558		
C_T	−0.528	0.630	0.886	
C_C	−0.450	0.487	0.619	0.584

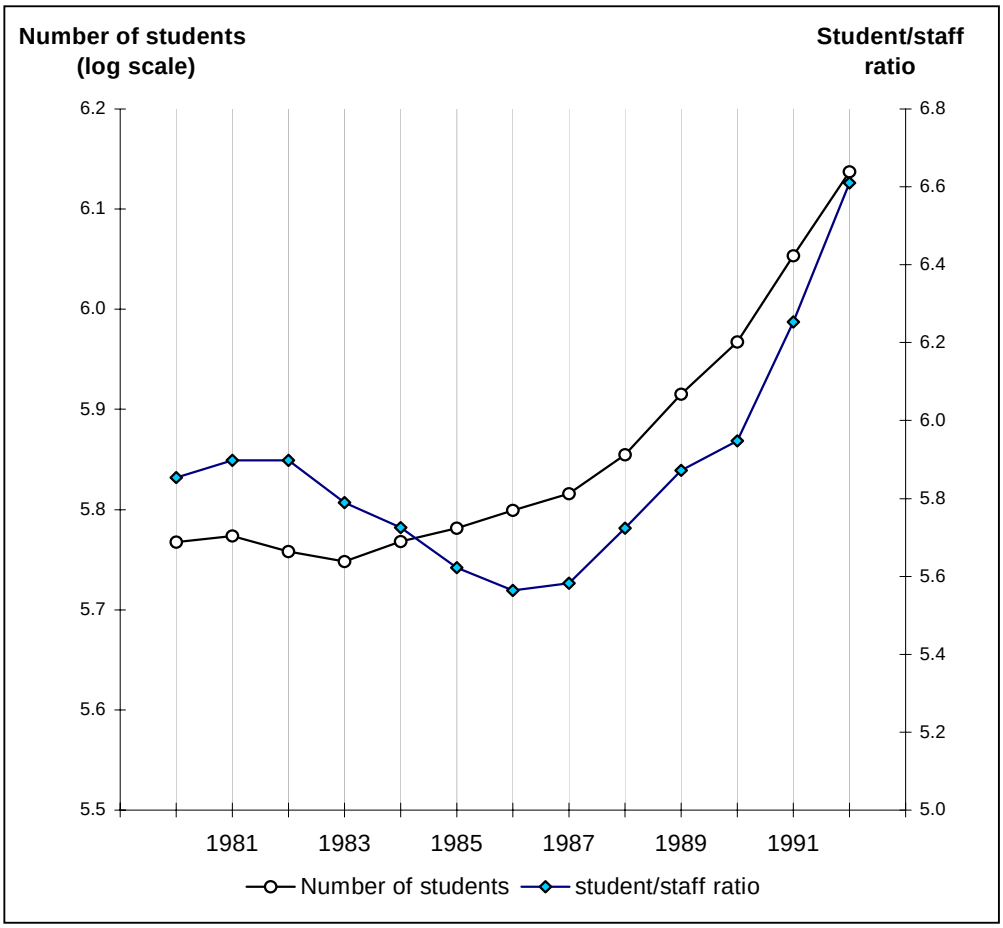

Fig. 1.

Students and staff: 45 UK universities, 1980/81–1992/93

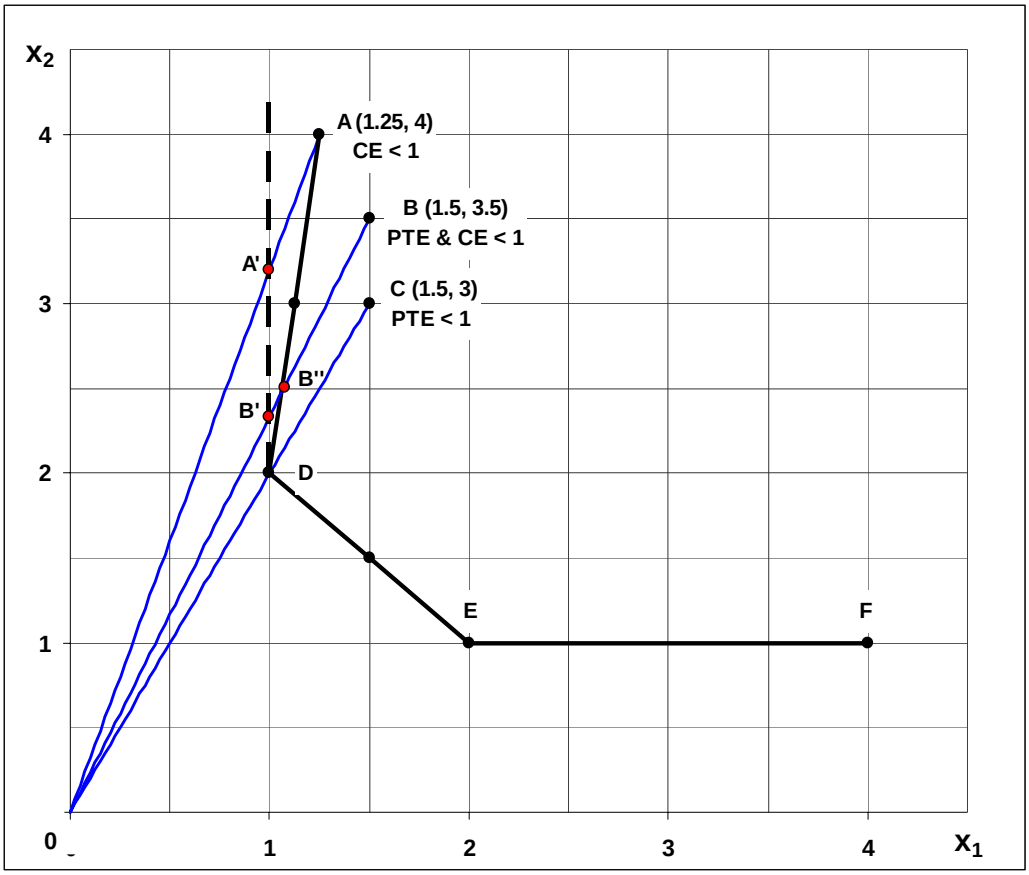

Fig. 2. Färe's approach (input-oriented, CRS)

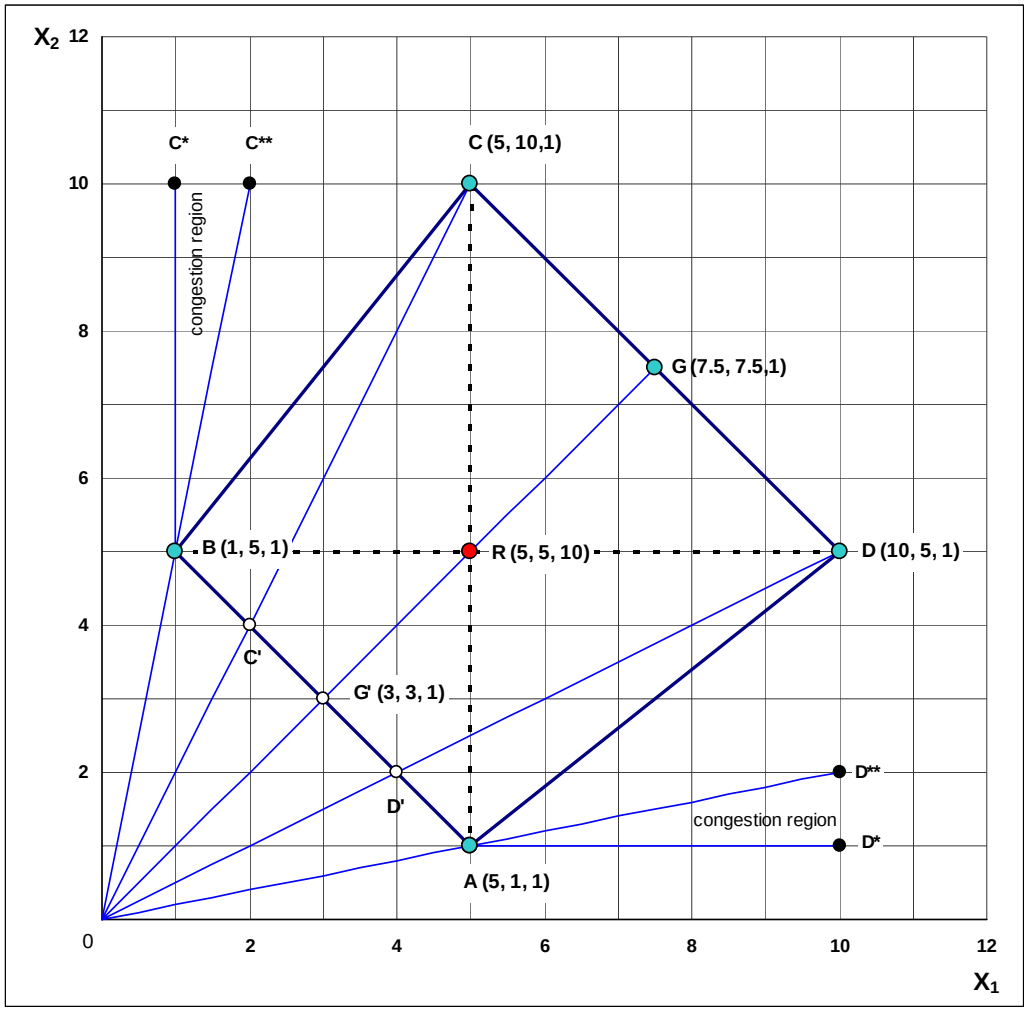

Fig. 3. A VRS model

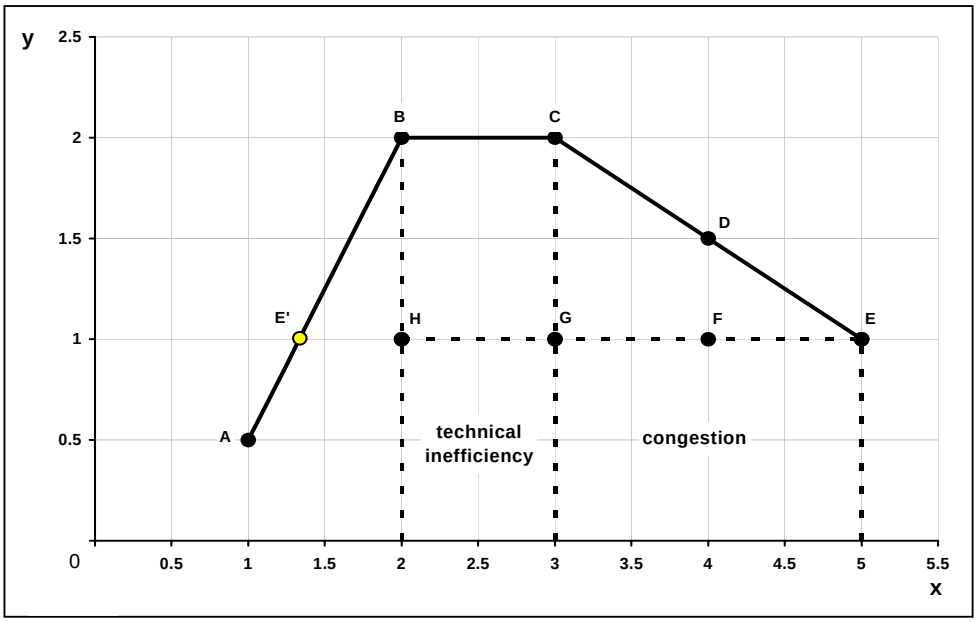

Fig. 4. Cooper's output-oriented approach

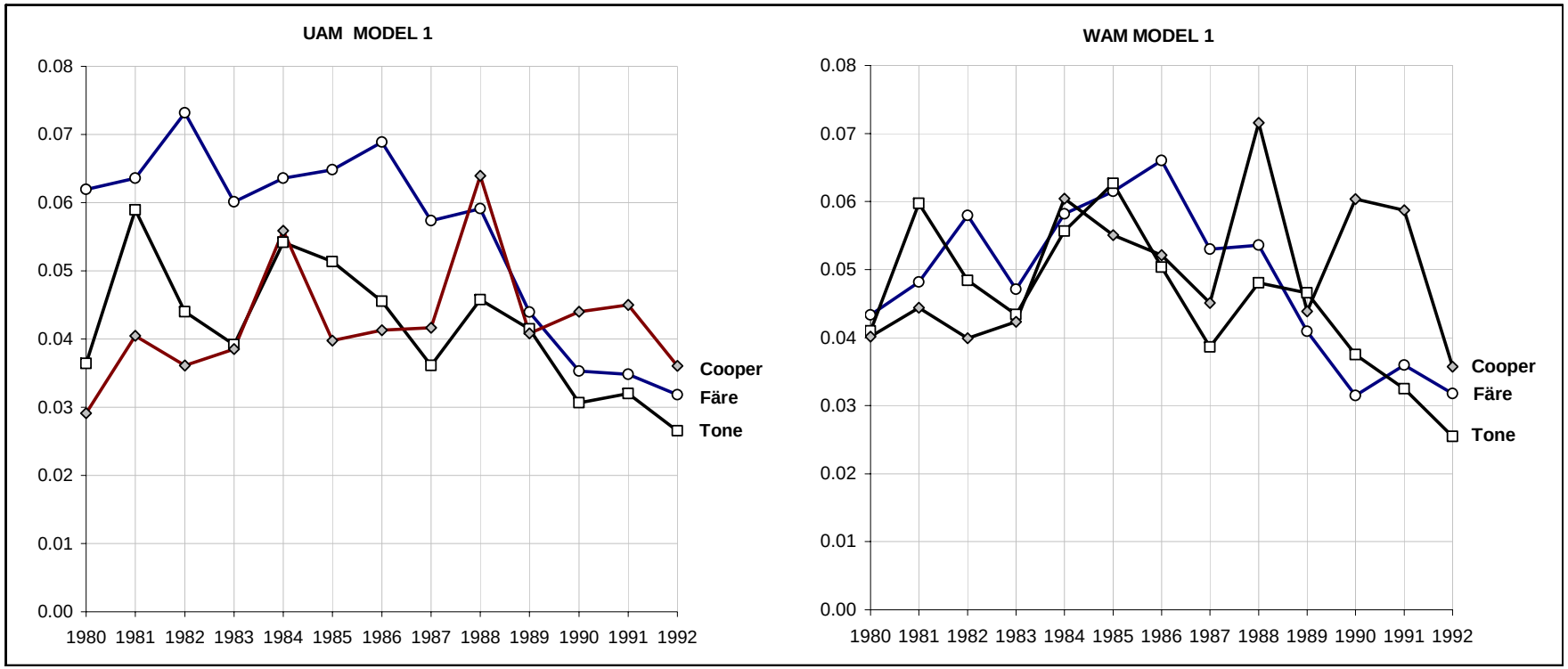

Fig. 5. Unweighted and weighted congestion scores for Model 1

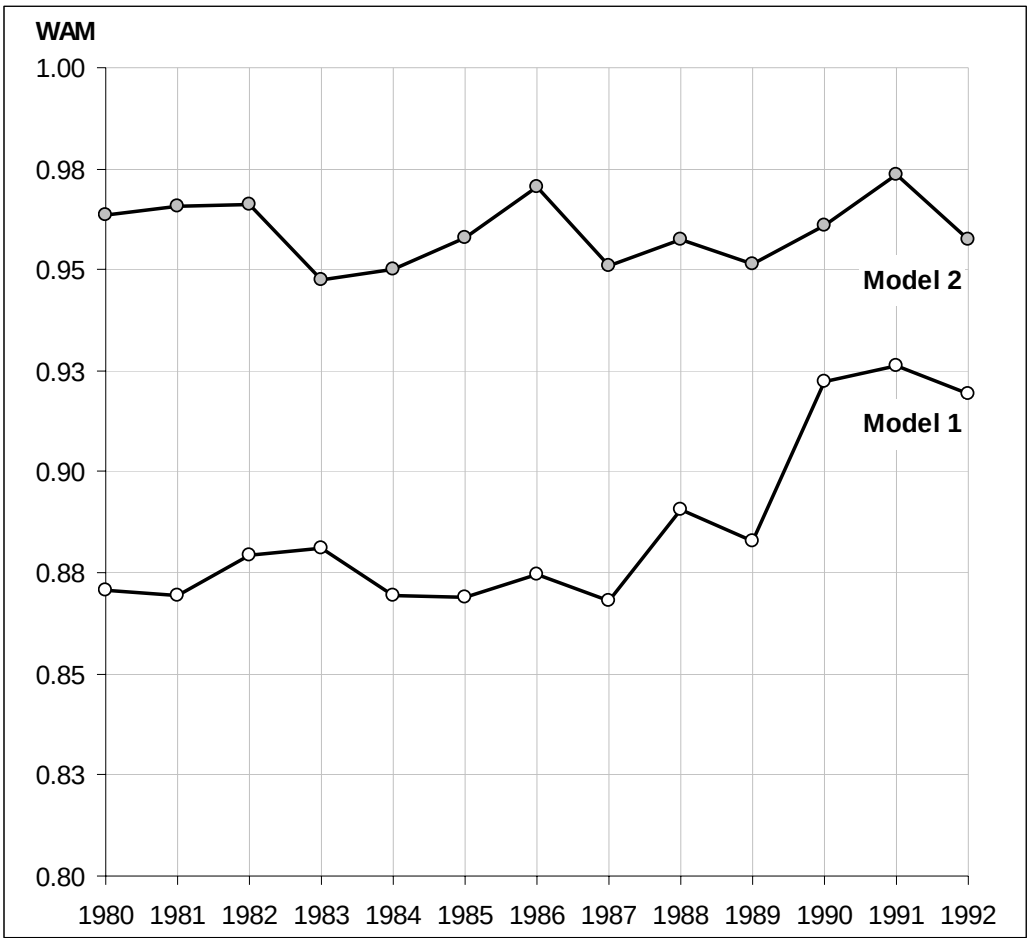

Fig. 6. Weighted mean TE scores for Models 1 and 2

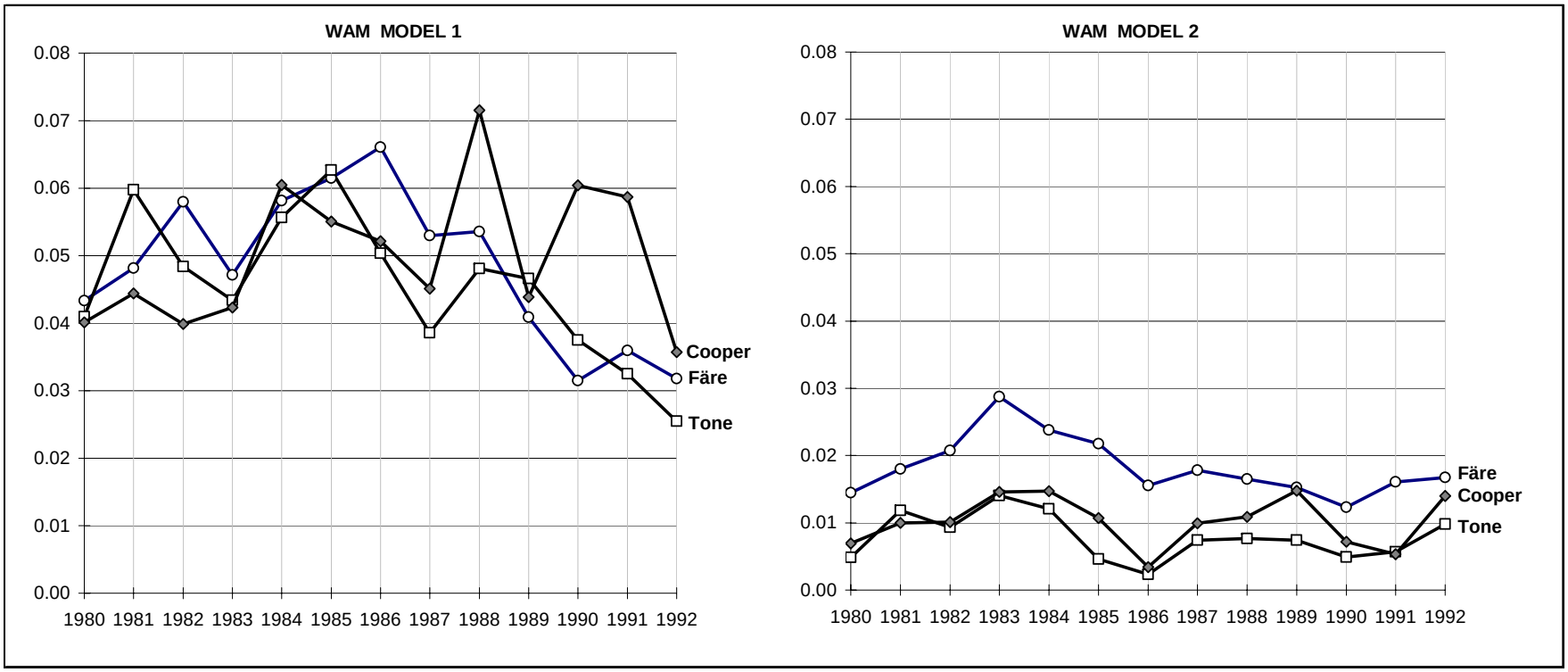

Fig. 7. Weighted congestion scores for Models 1 and 2

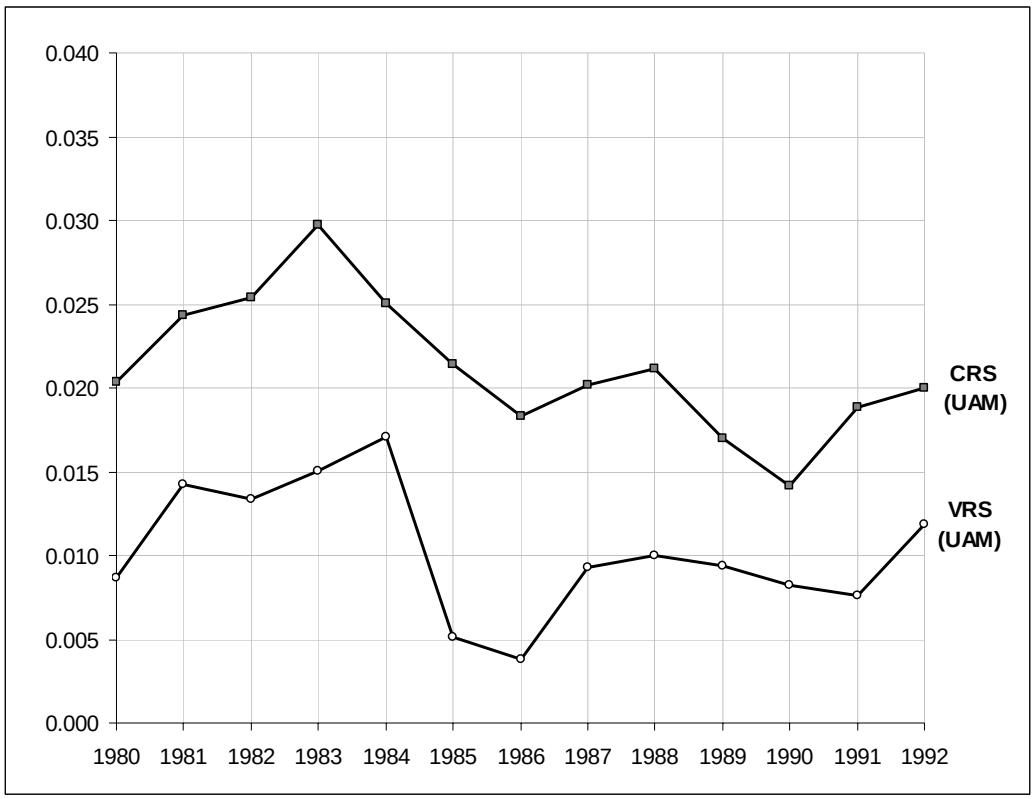

Fig. 8. Färe's congestion measure: CRS versus VRS (Model 2)

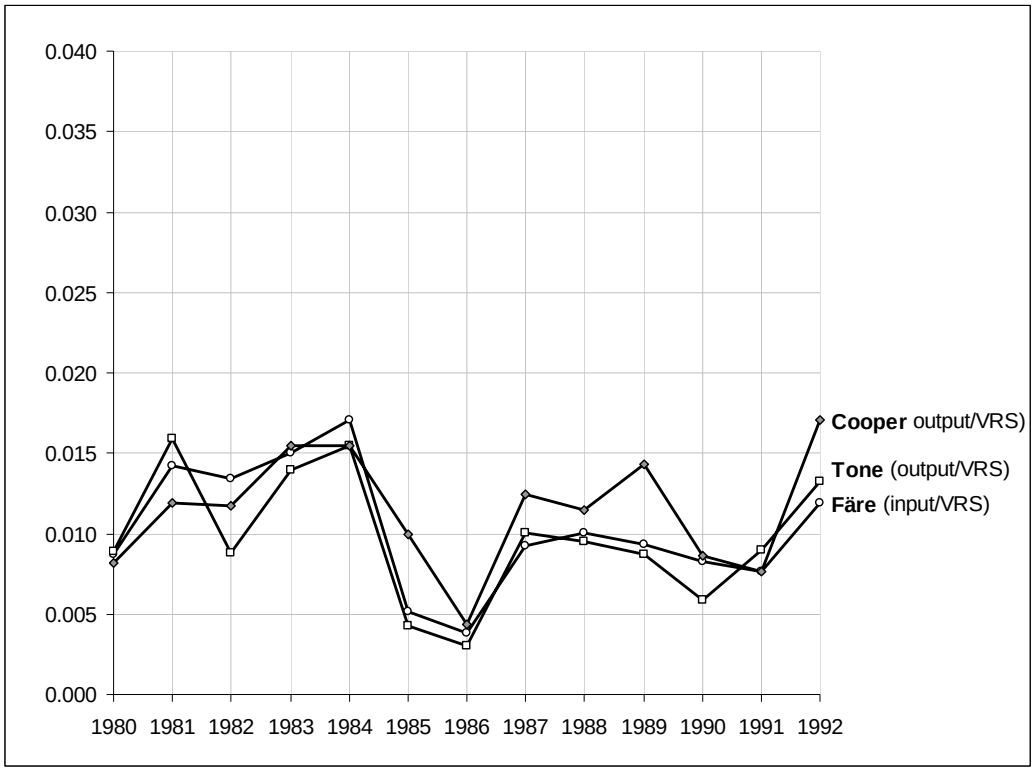

Fig. 9. The three measures of congestion compared (VRS, Model 2)