Does it matter how we measure congestion?

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 41 'new' British universities in the period 1995/6 to 2003/4. Contrary to expectations, Färe and Grosskopf's approach generally indicates substantially more congestion than do the other procedures. The main reason for this is identified as being its use of CRS rather than VRS as the assumed technology. While the alternative measures of congestion are found to be positively correlated, the correlations are not strong enough for them to be regarded as substitutes. All methods suggest the existence of a widespread problem of congestion in the new universities, although they differ noticeably as regards its severity.

KEY WORDS: Data envelopment analysis; Education; Congestion

1. Introduction

The focus of this paper is on the problem of congestion, which refers to a situation where the use of a particular input has increased by so much that output actually falls. Congestion can be viewed as an extreme form of technical inefficiency and, as such, can be regarded as a potentially serious practical problem. Consider, for instance, the case of universities. A substantial increase in the ratio of students to academic staff has been a common experience in universities throughout the world in recent decades. As a result, the marginal product of students might have become *negative* in some universities. The implication of this is that a

[*] 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.

reduction in the number of students, with all other inputs (staff, buildings, etc.) held constant, might raise a university's output in terms of research and degrees awarded, both undergraduate and postgraduate.

There has been much debate between the competing schools of thought about the appropriate way to measure congestion in the context of a DEA (data envelopment analysis) model, yet it seems fair to say that no consensus has been reached on the theoretical merits and demerits of the different approaches.[1] What is more, there is scant empirical evidence available as to whether the different approaches are likely to yield very different results.

This paper begins by comparing and contrasting the theoretical characteristics of the different approaches. This is done by examining hypothetical examples. A case study of British universities is then used to see whether the different approaches yield noticeably different results. This case study employs annual data relating to 41 former polytechnics that became universities in 1992. The analysis covers the period 1995/6 to 2003/4.

2. Defining 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 is indicated whenever more (less) of any input is employed, with all other inputs held constant, and there is a concomitant fall (rise) in output.* This

[1] For helpful discussions of these different approaches, see Cooper et al. (2004), Ray (2004, chapter 7) and Zhu (2002, chapter 9). Zhu also provides some useful software.

alternative definition is grounded in the hypothesis of diminishing marginal returns, with the added feature that congestion requires a negative marginal product to occur eventually.

3. Alternative approaches

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 by Cooper et al. (2000b, 2001c). For ease of exposition, these alternative procedures are referred to hereafter as the approaches of Färe and Cooper. More recently, Tone and Sahoo (2004) have proposed a new unified approach to measuring congestion and economies of scale. This new approach will also be examined in this paper.

The theoretical merits and demerits of the competing approaches of Cooper and Färe have been debated most recently by Cherchye et al. (2001) and Cooper et al. (2001a, 2001b), yet this debate was inconclusive. There is also little published information on whether these two approaches yield very different outcomes in terms of the measured amount of congestion. Hence it is important to consider carefully which approach or approaches to pursue.

4. Färe's approach

Figure 1 near here

Färe's approach is illustrated in Figure 1. Before examining this example, we should note that is possible to decompose Färe'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:

$$\text{TE} \equiv \text{PTE} \times \text{SE} \times \text{CE}, \tag{1}$$

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

Figure 1 shows six decision-making units (DMUs), each producing an output of y = 1, using two inputs, x_1 and x_2. This example assumes *constant returns to scale* (CRS), so that SE = 1, and makes use of an *input-oriented* approach. DMUs D and E are clearly technically efficient, whereas C is inefficient. In terms of identity (1) above, TE = PTE = ⅔ for C. Less obviously, F would also be deemed to be technically efficient under Färe's approach. Here the slack in x_1 of two units would be disregarded on the basis that these units were *freely disposable*, i.e. could be disposed of at no opportunity cost. Indeed, Färe and Grosskopf (2000a, pp. 32–33) argue that, given positive input prices, non-zero slack is akin to *allocative* rather than *technical* inefficiency.

The classification of DMUs A and B is both more complicated and more controversial. With respect to A, Färe's analysis would proceed along the following lines. Because A is on the isoquant for y = 1, Färe would regard this DMU as exhibiting no *pure* technical inefficiency (PTE = 1). However, it would be deemed to be suffering from congestion. A's CE score, as measured by the ratio OA´/OA, would equal 0.8. Its TE score would also equal 0.8, the product of PTE = 1 and CE = 0.8. According to Färe, congestion would arise owing to the difference between the upward-sloping isoquant segment DA, which is assumed to exhibit *weak* disposability, and the hypothetical vertical dashed line emanating from D, which is assumed to exhibit *strong* (or free) disposability. By moving to point A´, and thereby eliminating its congestion, A could attain TE = 1. By contrast, B would exhibit both pure technical inefficiency and congestion under Färe's approach. Here PTE = OB´´/OB ≈ 0.714 and CE = OB´/OB´´ ≈ 0.933, so that TE = ⅔ ≈ 0.714 × 0.933.[2]

However, Cooper would surely claim that there was no evidence that either A or B suffered from congestion![3] This is because all DMUs in Figure 1 produce the same output. For congestion to occur, in his view, one must observe a fall in output if the input in question

[2] The calculations were carried out using *OnFront*.

[3] The views attributed here to Cooper mirror those of Cooper et al. (2001a), who discuss an analogous example taken from Färe et al. (1985b, p. 76).

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.

In the context of this example, however, this particular criticism of Färe's approach seems 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.

5. Cooper's approach

At the outset, we need to define Cooper's measure of congestion, denoted here by C_C. The first step is to specify a formula for calculating the amount of congestion:

$$c_i = s_i^* - \delta_i^*, \tag{2}$$

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 (cf. Cooper et al., 2001a, p. 69). The asterisks denote optimal values generated by the DEA software. The measured amount of congestion is thus a residual derived from the DEA results. We can then rewrite equation (2) as follows:

$$c_i/x_i = s_i^*/x_i - \delta_i^*/x_i, \tag{3}$$

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 final step is to take arithmetic means over all inputs to get:[4]

$$C_C = \overline{s/x} - \overline{\delta/x}\,. \tag{4}$$

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).

Cooper's procedure makes use of the Banker–Charnes–Cooper (BCC) model, which assumes *variable returns to scale* (VRS). His procedure involves two steps. In the first step, the following *output-oriented* BCC model is employed to obtain the value of ϕ^* for each DMU k,[5] while the second step involves maximizing the sum of the slacks, conditional on this value of ϕ^* (cf. Cooper et al., 2000b, pp. 3–5):

$$\phi^* = \max \phi \tag{5a}$$

subject to:

$$\textstyle\sum_j \lambda_j x_{ij} \le x_{ik}, \qquad i = 1, 2, \ldots, m, \tag{5b}$$

$$\textstyle\sum_j \lambda_j y_{rj} \ge \phi y_{rk}, \qquad r = 1, 2, \ldots, s, \tag{5c}$$

$$\textstyle\sum_j \lambda_j = 1, \tag{5d}$$

$$\lambda_j \ge 0, \qquad j = 1, 2, \ldots, n. \tag{5e}$$

Figure 2 near here

The BCC model, in the context of a simplified production function y = f (x), is depicted in Figure 2 by the convex VRS frontier ABCDE and its horizontal extension from E. The diagram also shows, for comparison, the linear CRS frontier obtained from the

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

[5] Cooper et al. (2001d, p. 211) state that it would be wrong to use an input-oriented model when implementing their approach.

Charnes–Cooper–Rhodes (CCR) model, which is produced if we drop the constraint (5d).[6] The issue of congestion arises from the inclusion of DMUs F and G in the diagram.[7]

To illustrate the use of Cooper's model, consider DMU G in Figure 2. The diagram reveals that there are two possible referent DMUs available for evaluating G, viz D and E. Both would yield $\phi^* = 2.5$, yet D is the one that would maximize the slack in input x (giving $s_x = 3$ versus only 2 for E). Hence D 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 2, we would move along the BCC frontier from D to E, holding output constant at $y = 5$. This process would yield $\delta_x^* = 1$.

Thus, in the case of G, 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 $\overline{s/x} = 3/9$ and $\overline{\delta/x} = 1/9$, giving $C_C = 2/9 \approx 0.222$ for G. Likewise, for F, $C_C = (2/8 - 1/8) = 0.125$. As regards the other three DMUs, we would need to project them onto the frontier ABCDEFG. Their congestion status would then coincide with that of the projected DMU: $C_C = 0.125$ for H and $C_C = 0$ for I and J. E would be deemed to be technically inefficient but not congested. F would have $\phi^* = 5/4 = 1.25$, whereas G, H, I and J would have $\phi^* = 2.5$. Figure 2 also illustrates the point that the presence of slack is necessary but not sufficient for congestion to occur.

In reality, horizontal segments such as DE in Figure 2 are rare and, in the data set discussed later, no case occurs where non-zero slack exists, yet $\phi^* = 1$. If the BCC frontier does not have any DMUs like E, then the amount of congestion for each input normally equals the BCC slack for this input.[8] This greatly simplifies the work needed to compute C_C, as the second stage of Cooper's procedure is no longer required. Alternatively, one could use

[6] See Cooper et al. (2000a) for a detailed discussion of the CCR and BCC models.

[7] Cf. Tone and Sahoo (2004, Figure 2).

[8] We say *normally*, as it is also necessary to establish that all efficient DMUs are located at extreme points on the frontier, like A and B in Figure 2. This can be verified by running a 'super-efficiency' model; see Cooper et al. (2000b, pp. 15–17). Such models can be run using *DEA-Solver Pro*.

a variant of his approach, whereby the two stages are combined into a single model (Cooper et al., 2002). Unfortunately, this would entail sacrificing some useful information.

6. An illustrative example

Figure 3 near here

To clarify the differences between the approaches of Cooper and Färe, let us now consider Figure 3.[9] This shows six hypothetical DMUs, each producing a single output, y, using two inputs, x_1 and x_2. VRS is assumed. The figure takes the form of a pyramid with its pinnacle at M. Whereas M produces $y = 5$, the other five DMUs produce $y = 1$. M is clearly an efficient DMU but so too are A and B, regardless of whether we assume CRS or VRS.

Under Cooper's approach, DMUs C and D would be deemed to be congested. Both are located on upward-sloping isoquant segments; this arises because $MP_1 > 0$ and $MP_2 < 0$ along segment BC, whereas $MP_1 < 0$ and $MP_2 > 0$ along segment AD. Both DMUs have $C_C = 0.2$, calculated as ½{(0/6) + (4/10)} for C and ½{(4/10) + (0/6)} for D. The evaluation is relative to M in both cases.

E is an interesting case because it is located on a downward-sloping isoquant segment; this arises because $MP_1 < 0$ *and* $MP_2 < 0$. Here $C_C = ½\{(2/8) + (2/8)\} = 0.25$. The evaluation is again relative to M. Like C and D, E is deemed to be congested because a reduction in inputs is associated with a rise in output.

However, under Färe's approach, none of these three DMUs would be held to be congested! Instead, their inefficiency would be ascribed to the pure technical category. This finding can be explained by the fact that the projections onto the efficiency frontier occur along segment BA, at points C´, E´ and D´. In the identity TE ≡ PTE × SE × CE, TE = 0.2, PTE = 0.4375, SE ≈ 0.4571 and CE = 1 for all three DMUs.[10]

[9] A diagram similar to Figure 3 is the subject of a debate between Cherchye et al. (2001) and Cooper et al. (2001a, 2001b).

[10] This was confirmed using *OnFront* and an input-oriented model.

It is worth noting the circumstances in which a DMU *would* be found to be congested under Färe's approach. For instance, C would need to be repositioned at a point such as C*, so that the ray OC* intersected the vertical line emanating from point B. Likewise, D would need to be repositioned at a point such as D*, so that the ray OD* intersected the horizontal line emanating from point A.[11] This exercise illustrates the point that an upward-sloping isoquant (negative marginal product for *one* of the factors) is necessary but not sufficient for congestion to occur under Färe's approach. In fact, for congestion to be identified, the relevant isoquant segment would need to be relatively steep or relatively flat.

What would a relatively steep or relatively flat isoquant mean in economic terms? Since the gradient of an isoquant equals $-MP_1/MP_2$, any relatively flat isoquant segment (such as one joining points A and D* in Figure 3) would require a relatively small (negative) value for MP_1 but a relatively large (positive) value for MP_2. Similarly, any relatively steep isoquant segment (such as one joining points B and C* in Figure 3) would require a relatively small (negative) value for MP_2 but a relatively large (positive) value for MP_1. This analysis suggests that Färe's approach would tend to identify congestion where the factor in question had a marginal product that was only marginally negative (relative to the marginal product of the other factor) but fail to identify congestion where the marginal product was highly negative. This property seems counterintuitive.

DMU E is a rather different case: 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 rise in both x_1 and x_2 cannot reduce output. This eliminates 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.

[11] CE = Oc/OC* and CE = Od/OD* for the repositioned C and D, where CE = 0.8 in both cases.

What might be the underlying cause of congestion for a DMU like E? Cooper et al. (2001a, 2001b) do not examine this issue, although they criticize Färe's approach on the basis of its alleged adherence to the law of variable proportions. This 'law' can, in fact, be used to provide a rationale for the existence of congestion. First note that the region CDM is defined in terms of the equation $y = 17 - x_1 - x_2$, 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, 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 contravene Färe's axiom of weak disposability.

7. Merits and demerits of the two approaches

From the discussion in the previous section, it is clear that one should not expect the competing approaches of Cooper and Färe to yield the same outcomes in terms of congestion. It may be useful, therefore, to attempt to summarize the pros and cons of each approach.

For us, the most appealing aspect 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 (1). 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 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.
- Frontier DMUs (such as E in Figure 2) 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 E in Figure 3 does not exhibit congestion. Even so, many researchers – including the present authors – have used Färe's methodology to identify and measure congestion, so it is important that it should perform this additional task correctly too.

From our perspective, 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 apparently causing the problem and to what extent. By contrast, this information is more difficult to obtain from Färe's procedure (see Cooper et al., 2000b, pp. 6–7).[12] 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?

[12] To identify the congesting factors, one would need to run the model several times, each time making different assumptions about which inputs were 'strongly' or 'weakly' disposable. See Ray (2004, p. 183) for a discussion of this point.

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 et al. (2000b), 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.[13] In the present paper, we aim to add to the scant empirical evidence on this topic.

8. Congestion and diseconomies of scale

Tone and Sahoo (2004) have proposed a new unified approach to measuring congestion and scale economies. For simplicity, this procedure is referred to hereafter as Tone's approach. From our perspective, this approach has several attractive features. The first is that negative marginal productivity always signals congestion.[14] Secondly, the analysis can easily be done using the *DEA-Solver Pro* software. Thirdly, the output is comprehensive and easily understood. On the other hand, as with Cooper's approach, a straightforward decomposition of overall technical efficiency cannot be carried out.

Tone's approach is similar to that of Cooper inasmuch as a BCC output-oriented model is used initially, yet Tone measures congestion very differently. To explain his approach, let us return to the example in Figure 3.

[13] 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 CRS, panel data and three outputs.

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

Like Cooper, Tone would find A, B and M to be BCC efficient and hence not congested. The remaining DMUs would have a congestion score of $\psi = 5$, reflecting the fact that M is producing five times as much output as any of them. *DEA-Solver Pro* also provides us with a helpful figure for the *scale diseconomy*, ρ, for each congested DMU. For example, in the case of C, this is calculated as:

$$\rho = \frac{\%\text{ change in y}}{\%\text{ change in x}_1} = \frac{+400\%}{-40\%} = -10 \qquad (6)$$

Using the same method, we also get $\rho = -10$ for D. In the case of E, inputs fall by 25% on average, so that $\rho = -16$. These results suggest that congestion is equally serious for C and D but more serious for E. This finding is consistent with the outcome from Cooper's approach, where $C_C = 0.25$ for E but 0.2 for C and D. In Tone's terminology, we would describe E as being *strongly* congested (because both inputs are congested) but C and D as being *weakly* congested (because only one input is congested).

9. The case study

The case study employs annual data relating to 41 former British polytechnics. These institutions attained university status in 1992. The analysis covers the period 1995/6 to 2003/4. These new universities form a relatively homogeneous group, sharing a common history and facing similar opportunities and problems. In particular, they operate under much higher student : staff ratios than do the older British universities.[15] In addition, the older universities typically receive substantially more research funding per member of staff.

In view of this relative under-resourcing of these new universities, it seems worthwhile to investigate whether they are congested and, if so, whether this congestion has increased or decreased over time. Indeed, given the fact that the student : staff ratio has risen during the period under review, congestion may well have increased. A considerable advantage of

[15] The student : staff ratio in the ex-polytechnics was 17.5 in 1995/6 and 19.3 in 2003/4. The corresponding figures for the older universities were 7.5 and 9.4, respectively.

examining several years of data is that one can thereby avoid the possibility of the results being distorted by the use of an atypical year.

10. The model and methodology

Following previous research (see Flegg et al., 2004; Flegg and Allen, 2006, 2007), our DEA model presumes that a university's output can be measured by the benefits it provides in terms of teaching, research, consultancy and other educational services. These aspects of a university's activities are captured here via the following variables:

- income from research grants and contracts in £ thousands (y_1);
- the number of undergraduate qualifications awarded, adjusted for quality (y_2);
- the number of postgraduate degrees, diplomas and certificates awarded (y_3).

A detailed rationale for these variables is given in Flegg and Allen (2006), along with exact definitions and sources. Nonetheless, some discussion is required with regard to the second output variable. In Flegg et al. (2004), we employed a very narrow measure of undergraduate output, viz the number of first-class honours degrees. By contrast, in Flegg and Allen (2006), we formulated two alternative models: model 1 used the sum of first-class honours degrees and upper seconds, whereas model 2 used the sum of *all* undergraduate qualifications, including all degrees irrespective of classification, as well as diplomas and certificates. The latter type of output has become increasingly important in the new universities.[16] In this study, we have followed Johnes (2006), by constructing a weighted average of the various types of undergraduate award.

The undergraduate output variable, y_2, is defined as follows:

$$y_2 = 3 \times z_1 + 2.5 \times z_2 + 2 \times z_3 + 1.5 \times z_4 + z_5, \qquad (7)$$

[16] For the ex-polytechnics, 'other undergraduate awards' such as certificates and diplomas have gained in importance, rising from 27.7% of all undergraduate awards in 1995/6 to 34.3% in 2003/4.

where z_1 is the number of first-class honours degrees, z_2 is the number of upper seconds, z_3 is the number of lower seconds, z_4 is the number of third-class honours degrees, and z_5 is the sum of all other undergraduate qualifications, including unclassified and ‘pass’ degrees, as well as all undergraduate diplomas and certificates.[17] One reason for giving diplomas and certificates a lower weighting is that they normally involve a shorter period of study than do honours degrees.

Again following previous research, the resources used in producing the above-mentioned outputs are measured here via the following input variables:

- the number of full-time equivalent undergraduate students (x_1);
- the number of full-time equivalent postgraduate students (x_2);
- academic staff expenditure in £ thousands (x_3);
- other expenditure in £ thousands (x_4).

A rationale for these variables is provided in Flegg and Allen (2006), along with sources of data and other details.

In our earlier study of congestion in the older British universities in the period 1980/1 to 1992/3, we used an *output-oriented* variant of Färe’s approach 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 (see Flegg et al., 2004). Here we have modified our use of Färe’s approach to take into account 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.[18] 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

[17] A similar formula is used by Johnes (2006), although she does not appear to have included diplomas and certificates in the final category.

[18] See, for example, Byrnes et al. (1984) and Färe et al. (1985a).

which TE is decomposed.[19] Therefore, where congestion is anticipated on *a priori* grounds, Färe and Grosskopf recommend that one should specify CRS rather than VRS technology when measuring congestion. We have followed this suggestion here.

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.[20] 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 (cf. Färe et al., 1989). 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 expecting one or more of the inputs to be congested.

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 Figure 1. However, we will revisit this issue of the underlying technology later in the paper.

11. Mean congestion scores by method

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 all universities.[21] For consistency with Cooper's measure, the congestion efficiency (CE) scores from Färe's input-oriented approach were converted into *in*efficiency scores, viz $C_F \equiv 1 - \text{CE}$, before averaging over all universities. In the case of Tone's output-oriented approach, the following transformation was used: $C_T \equiv 1 - 1/\psi$, where

[19] 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.

[20] We are grateful to Pontus Roos, of the Institute of Applied Economics in Sweden, for clarifying this issue for us.

[21] n = 41 in the first seven years but 40 thereafter. This difference is due to a merger between two of the original polytechnics.

$\psi \geq 1$ is the congestion score generated by *DEA-Solver Pro*.[22] With these transformations, all measures have a convenient range from 0 (no congestion) to 1 (maximum congestion).

Table 1 & Figure 4 near here

The top panel of Table 1 shows the annual unweighted arithmetic mean (UAM) congestion scores for the three approaches and the corresponding rankings: F for Färe, T for Tone and C for Cooper. The bottom panel shows the results for the weighted arithmetic mean (WAM). Here the number of students in each university was used as a weight. The unweighted results, which are illustrated in Figure 4, will be examined first.

Figure 4 suggests that the period under review can be divided into two contrasting halves. During the first subperiod, there is evidence of a fall in congestion and all measures reach a minimum in 1999/0. Thereafter, all measures signal a rise in congestion, albeit by greatly differing amounts.

However, the measures do behave very differently: whereas $\overline{C}_C$ changes smoothly and consistently, the other two measures are much more erratic. 1996/7 is a case in point. This year witnessed a pronounced fall in the mean TE score, which is captured by the sharp rise in both $\overline{C}_F$ and $\overline{C}_T$. By contrast, Cooper's measure rises by a mere 0.0004! As noted later, there is a much weaker relationship between TE and congestion in the case of Cooper's approach than there is for the other two approaches.

An interesting facet of the results is that Färe's measure invariably signals more congestion than does Cooper's measure. Furthermore, for most years, there is a substantial gap between the respective graphs. In the light of the earlier discussion, this outcome is not what we had expected. As regards Tone and Cooper, the differences in mean scores are not so

[22] An alternative would be to define Tone's measure as $C_T \equiv \psi - 1$. Cooper et al. (2000b) 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.

marked, although Table 1 reveals that $\overline{C}_T > \overline{C}_C$ for eight years out of nine. What is more, taking the period as a whole, there is a fall in $\overline{C}_C$, yet little change in $\overline{C}_T$.

Whilst it is true that all three measures pick out 1996/7 as the year with the most congestion and 1999/0 as the year with the least, the differences in mean scores are large in the former case but small in the latter. Indeed, Figure 4 shows that there are only two years, viz 1997/8 and 1999/0, where there is a close correspondence between the three measures as regards the magnitude of congestion.

When the scores are weighted by the number of students in each university, a very similar picture emerges. In particular, as shown in Table 1, the earlier finding that $\overline{C}_F > \overline{C}_C$ is confirmed in all cases. Less clear-cut is the fact that $\overline{C}_T > \overline{C}_C$ for six (rather than eight) years out of nine. Some minor differences also appear when the results are averaged across methods and across years. However, what is most striking is the similarity of the weighted and unweighted results rather than the differences. This similarity is due to the fact that, with a few exceptions, the universities do not differ greatly in terms of size (see the Appendix). Therefore, for simplicity, only unweighted results will be discussed hereafter.

12. Scale diseconomies and congestion

Table 2 near here

Along with congestion scores, Tone's approach offers some useful information about diseconomies of scale. Table 2 shows the annual arithmetic mean values of ρ, Tone's *scale diseconomies* parameter, based on data for all universities. The table then shows the effect of excluding non-congested universities. Values of $\overline{C}_T$ are also displayed for comparison.

The results for all universities reveal that $\overline{C}_T$ and $\overline{\rho}$ yield very different rankings of years as regards the severity of congestion. For instance, whereas $\overline{C}_T$ ranks 1999/0 as the least congested year, $\overline{\rho}$ ranks it as the most congested! If we now look at the results for congested universities alone, it is evident that the values of ρ typically have a wide range and relatively high coefficient of variation (V). This variability is especially marked in the case of 1999/0, and this factor may well explain the conflicting rankings offered by $\overline{C}_T$ and $\overline{\rho}$.

In view of its sensitivity to extreme values, $\overline{\rho}$ is not a very reliable measure of the amount of congestion in a given year.[23] Nonetheless, the values of ρ do provide some very useful information about potential scale diseconomies in individual universities. Consider, for instance, the results for 2003/4, which are displayed in the Appendix. To take two extreme examples, these results suggest that a 1% decrease in congested inputs could have raised output in Manchester Metropolitan University by almost 25%, yet by only 1% in the University of the West of England, Bristol. However, 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.

13. Sources of Congestion

Table 3 near here

A useful attribute of Cooper's approach is that it is possible to assess, for each university, how much each input contributes to the observed amount of congestion. Table 3 takes a closer look at this facet of Cooper's method. The table reveals that, on average, excessive numbers of undergraduates (x_1) and postgraduates (x_2) account for almost half of the value of Cooper's congestion score, $\overline{C}_C$. However, the results also suggest that academic overstaffing is a major cause of congestion in the new British universities! Indeed, in six years out of nine, academic staff (x_3) account for a higher proportion of $\overline{C}_C$ than do undergraduates. Also rather surprising is the sizable role attributed to 'other expenditure' (x_4).

[23] Unlike C_T, ρ has no upper bound, and hence is likely to be more volatile as a result. It has a much larger coefficient of variation than C_T.

The finding regarding academic overstaffing is puzzling – especially in view of the high student : staff ratio in the new universities. 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, as well as undergraduate and postgraduate qualifications obtained. However, there is no obvious reason why this should occur, and it is possible that the presence of 'surplus' staff in the congested universities might indicate institutional inefficiency in a broader sense.

The role attributed to 'other expenditure' is equally puzzling. What this suggests is that, beyond a certain point, extra expenditure actually reduced congested universities' output. However, a possible explanation is in terms of differences in the *mix* of expenditure in different universities. 'Other expenditure' is a very broadly defined input variable, comprising expenditure on academic cost centres, academic services, administration and central services, premises, residences and catering, and on research grants and contracts. It is conceivable, for instance, that a high proportion of 'other expenditure' devoted to research could impact adversely on the output of undergraduate qualifications, even though it might stimulate research output. Another possible explanation is in terms of excessive spending on administration, which could reduce a university's efficiency and hence output in terms of research and qualifications awarded.

14. Order of Decomposition

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. To explore this issue, we recalculated C_F using VRS.[24] The results are presented in Table 4 and illustrated in Figure 5.

Table 4 & Figure 5 near here

Figure 5 reveals that, for most years, we get appreciably less ‘congestion’ if we assume VRS rather than CRS. This can be confirmed by comparing the columns headed $\overline{C}_{F,CRS}$ and $\overline{C}_{F,VRS}$ in Table 4. What is more, the table shows that there is little difference between Färe’s VRS-based measure and that of Tone.

Of the three VRS-based methods, Cooper’s method stands out as being the most different. Table 4 also confirms the earlier finding that it tends to indicate the least congestion. However, while it is true that $\overline{C}_{F,VRS} > \overline{C}_{C}$ for eight of the nine years, it is noticeable how the gap between $\overline{C}_{F,VRS}$ and $\overline{C}_{C}$ is usually smaller than that between $\overline{C}_{F,CRS}$ and $\overline{C}_{C}$.

Table 5 near here

To shed some more light on the relationships among the different measures, correlation coefficients were calculated using the raw congestion scores (n = 367). Table 5 shows the results. As anticipated, $C_{F,VRS}$ is very strongly correlated with C_T. The fact that this correlation is 0.898 rather than unity can be attributed to the different orientation and to the different ways in which congestion is measured.

C_T is also strongly correlated with $C_{F,CRS}$. This result was not anticipated but it reflects the fact that Färe’s two measures are themselves fairly strongly correlated (r = 0.789). Table 5 also shows that Cooper’s measure is not strongly correlated with any of the other three measures. As expected, all measures are negatively correlated with TE; this finding

[24] 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). In his discussion of Färe’s approach, Ray (2004, pp. 170, 175−86) employs VRS models throughout; he does not raise the issue of whether one should use VRS or CRS technology.

suggests that a fall in congestion would raise technical efficiency. However, the correlation is rather weak in the case of C_C.

The correlation analysis shows that the four measures are positively associated, yet the strength of this correlation varies substantially and some measures appear to be more substitutable than others. Even so, the correlations need to be interpreted with care. For instance, $C_{F,CRS}$ is strongly correlated with C_T (r = 0.825), yet $C_{F,CRS}$ is apt to identify a lot more congestion than C_T would do. More detailed information is given in the Appendix, where individual results for 2003/4 are tabulated.

There is, in fact, a very close correspondence between the sets of universities deemed to be congested by the three VRS-based methods. For instance, the Appendix shows that they identify the same 21 universities as being congested in 2003/4, whereas Färe's CRS-based procedure uncovers an extra seven congested universities. Similar results were found for the other eight years.[25]

This close matching of the universities deemed to be congested by the VRS-based methods is a little surprising at first sight. However, in the case of Cooper and Tone, it can be explained by the fact 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.[26]

[25] Over the period as a whole, there were only five cases out of 367 where Cooper and Tone would disagree as to whether a given university was or was not congested. Likewise, there were only six instances where $C_{F,VRS}$ and C_C gave conflicting results. By contrast, $C_{F,CRS}$ identified 245 cases of congestion, whereas $C_{F,VRS}$ found only 180 cases.

[26] Tone uses an output-oriented version of the slacks-based measure (Tone, 2001) to project each congested DMU onto the BCC frontier.

It is harder to explain why Färe's VRS-based measure should identify the same set of congested universities, as some differences were anticipated owing to the different orientation and the fact that his method employs a radial projection.

The fact that Färe's CRS-based procedure identifies an extra seven congested universities in 2003/4 is worth exploring. These universities are Abertay Dundee, Central England, London Metropolitan, Luton, Northumbria, Paisley and Robert Gordon. As shown in the Appendix, Färe's CRS-based procedure attributes the technical inefficiency of these universities entirely to congestion, whereas his VRS-based procedure indicates a complete absence of congestion! The BCC model (which assumes VRS) ascribes the technical inefficiency of these universities wholly to an inappropriate scale. This is shown by the fact that TE = SE in all seven cases.[27] Therefore, whether these seven universities are deemed to exhibit scale inefficiency or congestion depends crucially on what assumption one makes about the underlying technology.

15. Conclusion

This paper has examined three alternative approaches to measuring congestion: the conventional approach of Färe and Grosskopf, the alternative proposed by Cooper et al., and a new procedure developed by Tone and Sahoo. In addition, two versions of Färe and Grosskopf's approach were considered: one assumed constant returns to scale (CRS), while the other assumed variable returns (VRS). At the outset, the methods were examined using hypothetical examples. The aim here was to highlight the theoretical properties of the different measures. This was followed by a case study of 41 former British polytechnics that became universities in 1992. This case study employed annual data for the period 1995/6 to 2003/4.

The four alternative methods indicated differing amounts of congestion, although Tone and Sahoo's method and the VRS-based version of Färe and Grosskopf's approach generated the

[27] SE is calculated as the ratio of the CCR and BCC efficiency scores (e.g., SE = 0.5 for DMU A in Figure 2).

most similar congestion scores. For instance, in 2003/4, the former indicated congestion of 5.2%, on average, across the 40 universities, whereas the latter indicated 5.4%. When the scores were averaged over the 21 congested universities, the figures were still similar, albeit much higher, viz 9.9% and 10.3%, respectively. Cooper's method generated the lowest average scores of the four methods: 2.4% for the whole sample and 4.5% for the congested universities.

Switching from VRS to CRS had a marked impact on the results from Färe and Grosskopf's approach: the mean congestion scores were substantially higher in almost all years. What is more, this method consistently produced the highest congestion scores of the four methods examined here. For instance, the mean score for the whole sample was 7.0% in 2003/4, well above the 5.4% for the VRS-based variant of their procedure, the 5.2% for Tone and Sahoo's method and the 2.4% for Cooper's method.

It is worth noting too that, on several occasions, the four measures exhibited rather different trends during the period under review. For instance, while all measures reached a minimum in 1999/0, Färe's CRS-based measure pointed to rising congestion thereafter, whereas Cooper's measure indicated negligible change!

Thus it *does* matter how congestion is measured. Since the different methods all have their respective theoretical merits and demerits, yet produce different results, it would seem sensible not to rely on a single method. For the same reason, relying upon the rankings generated by a single method would be unwise. However, if one's aim is simply to classify universities into sets of congested and uncongested institutions, then it makes little difference which of the three VRS-based methods is employed, although it does make a great deal of difference whether one opts for CRS or VRS technology.

The choice of technology is clearly an important issue: if we follow Färe and Grosskopf in positing CRS technology in cases where congestion is anticipated on *a priori* grounds, then we are likely to find more congestion and rather less scale inefficiency. This matters because the remedies for the two types of inefficiency are apt to be very different.

There are several factors that lend credence to the results obtained here. The first is that the study examined data for nine years, thereby minimizing the possibility of the results being influenced unduly by the peculiarities of particular years. The relevance of this point is illustrated by the fact that the four methods generated similar results in 1997/8 and 1999/0, yet very different results in the other seven years! The second point is that the results were not materially affected by changes in the definitions of the variables used in the DEA models (for details, see Flegg and Allen, 2006). The final point is that we obtained broadly similar results in our earlier study of 45 traditional British universities over the same period (see Flegg and Allen, 2007), although here it is worth noting that Cooper's method typically generated *more* congestion than did the VRS-based variant of Färe and Grosskopf's approach. This suggests that there is no general empirical relationship between these two methods. This is unsurprising, given the very different ways in which congestion is measured.

In terms of future work, it would be worthwhile to investigate the reasons why Cooper's decomposition analysis should ascribe such a large role to academic overstaffing. Here it might be fruitful to make use of the facility in *OnFront*, whereby one can restrict consideration to a subset of inputs most likely to be affected by congestion.

References

Brockett PL, Cooper WW, Shin HC, Wang Y, 1998. Inefficiency and congestion in Chinese production before and after the 1978 economic reforms. Socio-Economic Planning Sciences 32; 1–20.

Byrnes P, Färe R, Grosskopf S, 1984. Measuring productive efficiency: an application to Illinois strip mines. Management Science 30; 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; 75–80.

Cooper WW, Deng H, Gu B, Li S, Thrall RM, 2001c. Using DEA to improve the management of congestion in Chinese industries (1981−1997). Socio-Economic Planning Sciences 35; 227–242.

Cooper WW, Deng H, Huang ZM, Li SX, 2002. A one-model approach to congestion in data envelopment analysis. Socio-Economic Planning Sciences 36; 231–238.

Cooper WW, Deng H, Seiford LM, Zhu J, 2004. Congestion: its identification and management with DEA. In: Cooper WW, Seiford LM, Zhu J. (Eds), Handbook of Data Envelopment Analysis. Kluwer: Boston; pp. 177–201.

Cooper WW, Gu B, Li S, 2001a. Comparisons and evaluations of alternative approaches to the treatment of congestion in DEA. European Journal of Operational Research 132; 62–74.

Cooper WW, 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; 81–87.

Cooper WW, Seiford LM, Tone K, 2000a. Data Envelopment Analysis. Kluwer: Boston.

Cooper WW, Seiford LM, Zhu J, 2000b. A unified additive model approach for evaluating inefficiency and congestion with associated measures in DEA. Socio-Economic Planning Sciences 34; 1–25.

Cooper WW, Seiford LM, Zhu J, 2001d. Slacks and congestion: response to a comment by R. Färe and S. Grosskopf. Socio-Economic Planning Sciences 35; 205–215.

Cooper WW, Thompson RG, Thrall RM, 1996. Introduction: extensions and new developments in DEA. Annals of Operations Research 66; 3–45.

Färe R, Grosskopf S, 2000a. Slacks and congestion: a comment. Socio-Economic Planning Sciences 34; 27–33.

Färe R, Grosskopf S, 2000b. Research note: decomposing technical efficiency with care. Management Science 46; 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; 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; 89–106.

Färe R, Grosskopf S, Lovell CAK, 1985b. The Measurement of Efficiency of Production. Kluwer-Nijhoff: Boston.

Färe R, Grosskopf S, Lovell CAK, Pasurka C, 1989. Multilateral productivity comparisons when some outputs are undesirable: a non-parametric approach. Review of Economics and Statistics 71; 90–98.

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

Flegg AT, Allen DO, Field K, Thurlow, TW, 2004. Measuring the efficiency of British universities: a multi-period data envelopment analysis. Education Economics 12; 231–249.

Flegg AT, Allen DO, 2006. Are the new British universities congested? Discussion Paper 06/10, School of Economics, Bristol Business School, University of the West of England, Bristol. http://carecon.org.uk/DPs/. Paper presented at the 48th Annual Conference of the Operational Research Society, Bath, September 2006.

Flegg AT, Allen DO, 2007. Does expansion cause congestion? The case of the older British universities, 1994 to 2004. Education Economics 15; forthcoming.

Johnes J, 2006. Data envelopment analysis and its application to the measurement of efficiency in higher education. Economics of Education Review 25; 273–288.

Ray SC, 2004. Data Envelopment Analysis: Theory and Techniques for Economics and Operations Research. Cambridge University Press: Cambridge.

Tone K, 2001. A slacks-based measure of efficiency in data envelopment analysis. European Journal of Operational Research 130; 498–509.

Tone K, Sahoo BK, 2004. Degree of scale economies and congestion: a unified DEA approach. European Journal of Operational Research 158; 755–772.

Zhu J, 2002. Quantitative Models for Performance Evaluation and Benchmarking: Data Envelopment Analysis with Spreadsheets. Kluwer: Boston.

Table 1. Alternative measures of congestion (all universities)

	Unweighted Arithmetic Mean (UAM)				**Ranking by method**			**Ranking of year**		
	Färe	**Tone**	**Cooper**	**Mean**	**F**	**T**	**C**	**F**	**T**	**C**
1995/6	0.0884	0.0489	0.0483	0.0619	1	2	3	8	7	8
1996/7	0.1425	0.0819	0.0487	0.0910	1	2	3	**9**	**9**	**9**
1997/8	0.0505	0.0385	0.0443	0.0444	1	3	2	4	5	7
1998/9	0.0542	0.0386	0.0352	0.0427	1	2	3	6	6	6
1999/0	0.0272	0.0240	0.0197	0.0230	1	2	3	**1**	**1**	**1**
2000/1	0.0347	0.0371	0.0210	0.0309	2	1	3	2	4	2
2001/2	0.0452	0.0334	0.0222	0.0336	1	2	3	3	3	3
2002/3	0.0520	0.0329	0.0238	0.0362	1	2	3	5	2	5
2003/4	0.0696	0.0518	0.0236	0.0483	1	2	3	7	8	4
Min	0.0272	0.0240	0.0197	0.0230						
Max	0.1425	0.0819	0.0487	0.0910						
Mean	0.0627	0.0430	0.0319	0.0458						
SD	0.0349	0.0168	0.0123	0.0203						

	Weighted Arithmetic Mean (WAM)				**Ranking by method**			**Ranking of year**		
	Färe	**Tone**	**Cooper**	**Mean**	**F**	**T**	**C**	**F**	**T**	**C**
1995/6	0.0832	0.0443	0.0475	0.0583	1	3	2	8	7	8
1996/7	0.1385	0.0783	0.0505	0.0891	1	2	3	**9**	**9**	**9**
1997/8	0.0520	0.0405	0.0471	0.0465	1	3	2	6	6	7
1998/9	0.0509	0.0364	0.0375	0.0416	1	3	2	5	5	6
1999/0	0.0242	0.0225	0.0207	0.0225	1	2	3	**1**	**1**	**1**
2000/1	0.0307	0.0314	0.0210	0.0277	2	1	3	2	2	2
2001/2	0.0392	0.0348	0.0221	0.0320	1	2	3	3	4	3
2002/3	0.0498	0.0327	0.0247	0.0357	1	2	3	4	3	4
2003/4	0.0628	0.0484	0.0247	0.0453	1	2	3	7	8	5
Min	0.0242	0.0225	0.0207	0.0225						
Max	0.1385	0.0783	0.0505	0.0891						
Mean	0.0590	0.0410	0.0329	0.0443						
SD	0.0345	0.0159	0.0127	0.0200						

Table 2. Scale diseconomies and congestion (unweighted): Tone’s approach

	All universities				Congested universities					
	$\overline{C}_T$	Rank	$\overline{\rho}$	Rank	Number	$\overline{C}_T$	$\overline{\rho}$	Max	Min	V
1995/6	0.0489	7	–4.60	6	24	0.0835	–7.86	–83.0	–0.14	16.7
1996/7	0.0819	**9**	–5.25	7	24	0.1399	–8.97	–31.1	–0.64	8.1
1997/8	0.0385	5	–2.03	2	21	0.0751	–3.97	–8.8	–0.87	2.3
1998/9	0.0386	6	–4.39	5	18	0.0879	–9.99	–81.1	–0.78	18.5
1999/0	0.0240	**1**	–7.46	**9**	19	0.0518	–16.10	–211.4	–0.84	46.2
2000/1	0.0371	4	–5.31	8	17	0.0894	–12.81	–107.9	–0.26	24.5
2001/2	0.0334	3	–2.03	**1**	17	0.0804	–4.89	–16.8	–0.88	4.1
2002/3	0.0329	2	–3.96	3	18	0.0732	–8.80	–56.6	–0.39	14.9
2003/4	0.0518	8	–4.12	4	21	0.0986	–7.85	–24.7	–0.31	7.3

Table 3. Percentage contribution of each input to congestion in congested universities: Cooper’s approach

	Other expenditure	Academic staff	Postgrads	Undergrads	Number congested	$\overline{C}_C$ (UAM)
1995/6	15.2	38.1	18.3	28.5	24	0.0825
1996/7	17.8	30.2	14.7	37.3	26	0.0768
1997/8	15.9	31.4	35.4	17.3	21	0.0864
1998/9	2.9	40.5	39.6	17.0	19	0.0760
1999/0	21.4	42.2	19.4	16.9	19	0.0424
2000/1	13.5	32.0	29.3	25.2	18	0.0478
2001/2	27.2	20.7	31.1	21.0	18	0.0505
2002/3	26.2	26.4	20.3	27.1	18	0.0529
2003/4	19.4	32.1	23.5	25.1	21	0.0449
Mean	17.7	32.6	25.7	23.9		0.0687

Table 4. Results from different approaches: Färe versus Cooper and Tone (unweighted, all universities)

	$\overline{C}_{F,VRS}$	$\overline{C}_{F,VRS}-\overline{C}_T$	$\overline{C}_{F,VRS}-\overline{C}_C$	$\overline{C}_{F,CRS}$	$\overline{C}_{F,CRS}-\overline{C}_T$	$\overline{C}_{F,CRS}-\overline{C}_C$
1995/6	0.0641	0.0152	0.0158	0.0884	0.0395	0.0401
1996/7	0.0879	0.0060	0.0392	0.1425	0.0606	0.0938
1997/8	0.0424	0.0039	–0.0019	0.0505	0.0120	0.0062
1998/9	0.0354	–0.0032	0.0002	0.0542	0.0156	0.0189
1999/0	0.0270	0.0030	0.0074	0.0272	0.0032	0.0075
2000/1	0.0380	0.0010	0.0170	0.0347	–0.0024	0.0137
2001/2	0.0310	–0.0023	0.0089	0.0452	0.0119	0.0231
2002/3	0.0305	–0.0025	0.0066	0.0520	0.0191	0.0282
2003/4	0.0541	0.0024	0.0306	0.0696	0.0179	0.0460
Mean	0.0456	0.0026	0.0137	0.0627	0.0197	0.0308

Table 5. Correlations: n = 367

	TE	C_T	C_C	$C_{F, CRS}$
C_T	–0.666			
C_C	–0.414	0.542		
$C_{F, CRS}$	–0.711	0.825	0.531	
$C_{F, VRS}$	–0.727	0.898	0.539	0.789

Appendix. Individual results for 2003/4

						Färe				Tone			Cooper	
University	Weight	TE	RANK	SE	RANK	$C_{F,CRS}$	RANK	$C_{F,VRS}$	RANK	C_T	RANK	ρ	C_C	RANK
Abertay Dundee	0.007	0.8534	23	0.8534	1	0.1466	34	0	1	0	1		0	1
Anglia Polytechnic	0.028	0.9144	16	0.9658	21	0.0856	28	0.0583	26	0.0532	29	−0.91	0.0232	26
Bournemouth	0.020	0.8246	30	0.9781	32	0.0776	27	0.1131	32	0.0496	28	−6.14	0.0373	29
Brighton	0.024	0.8761	18	0.9787	26	0.0040	13	0.0020	20	0.0024	20	−0.31	0.0217	25
Central England	0.029	0.9634	13	0.9634	1	0.0366	21	0	1	0	1		0	1
Central Lancashire	0.033	1	1	1	1	0	1	0	1	0	1		0	1
Coventry	0.022	0.8489	25	0.9559	27	0.1511	36	0.1241	34	0.1119	33	−9.34	0.0182	24
De Montfort	0.030	1	1	1	1	0	1	0	1	0	1		0	1
Derby	0.018	0.8213	31	0.9669	31	0.1081	31	0.1204	33	0.0097	23	−7.02	0.0528	34
East London	0.019	0.8528	24	0.9996	30	0.1472	35	0.1057	30	0.1469	35	−2.22	0.0975	38
Glamorgan	0.022	0.7707	33	0.9686	35	0.043	22	0.1624	36	0.2043	36	−6.74	0.0121	22
Glasgow Caledonian	0.022	0.6798	40	0.9181	39	0.3202	40	0.1917	38	0.2596	39	−3.89	0.1199	40
Greenwich	0.025	1	1	1	1	0	1	0	1	0	1		0	1
Hertfordshire	0.030	0.7619	36	0.9806	36	0.0269	18	0.0191	21	0.0295	26	−4.64	0.0157	23
Huddersfield	0.021	1	1	1	1	0	1	0	1	0	1		0	1
Kingston	0.027	0.7364	37	0.9128	34	0.0226	17	0.0819	27	0.0037	21	−4.31	0.0571	36
Leeds Metropolitan	0.034	0.8543	22	0.9115	22	0.0151	15	0.0896	28	0.0627	30	−15.61	0.0351	27
Lincoln	0.017	0.8399	27	0.9687	28	0.0126	14	0.1051	29	0.133	34	−24.28	0.0353	28
Liverpool J. Moores	0.027	1	1	1	1	0	1	0	1	0	1		0	1
London Metro	0.036	0.8396	28	0.8396	1	0.0986	29	0	1	0	1		0	1
London South Bank	0.022	1	1	1	1	0	1	0	1	0	1		0	1
Luton	0.013	0.7932	32	0.7932	1	0.0591	25	0	1	0	1		0	1
Manchester Metro	0.045	0.9243	15	0.9511	20	0.0757	26	0.0308	24	0.0282	25	−24.74	0.1077	39
Middlesex	0.027	1	1	1	1	0	1	0	1	0	1		0	1
Napier	0.015	0.7314	38	0.9766	38	0.2686	38	0.2114	39	0.2511	38	−12.91	0.0433	32
Northumbria	0.032	0.9498	14	0.9498	1	0.0502	23	0	1	0	1		0	1
Nottingham Trent	0.039	1	1	1	1	0	1	0	1	0	1		0	1
Oxford Brookes	0.023	0.8436	26	0.9326	24	0.0562	24	0.1114	31	0.0954	32	−1.34	0.0552	35
Paisley	0.013	0.8683	19	0.8683	1	0.1317	32	0	1	0	1		0	1
Plymouth	0.034	0.8371	29	0.9806	29	0.0299	19	0.0399	25	0.0063	22	−4.18	0.0106	21
Portsmouth	0.028	1	1	1	1	0	1	0	1	0	1		0	1
Robert Gordon	0.014	0.9004	17	0.9004	1	0.0996	30	0	1	0	1		0	1
Sheffield Hallam	0.037	0.7683	34	0.9923	37	0.2317	37	0.1829	37	0.2258	37	−9.51	0.0411	30
Staffordshire	0.019	1	1	1	1	0	1	0	1	0	1		0	1
Sunderland	0.019	1	1	1	1	0	1	0	1	0	1		0	1
Teesside	0.021	1	1	1	1	0	1	0	1	0	1		0	1
Thames Valley	0.019	0.7095	39	0.9630	40	0.2905	39	0.2323	40	0.2632	40	−19.82	0.0445	33
West of England	0.037	0.7679	35	0.9181	33	0.0218	16	0.0267	23	0.0389	27	−1.00	0.0425	31
Westminster	0.026	0.8620	20	0.9385	23	0.1380	33	0.1340	35	0.0815	31	−1.94	0.0692	37
Wolverhampton	0.028	0.8574	21	0.9504	25	0.0355	20	0.0224	22	0.0134	24	−4.06	0.0029	20
Mean	0.025	0.8813		0.9569		0.0696		0.0541		0.0518		−4.12	0.0236	
Number on frontier		12		12		12		19		19			19	
Correlations: TE						−0.6780		−0.7354		−0.6743			−0.5612	
$C_{F,\,CRS}$								0.7584		0.7797			0.5568	
$C_{F,\,VRS}$										0.9222			0.6500	

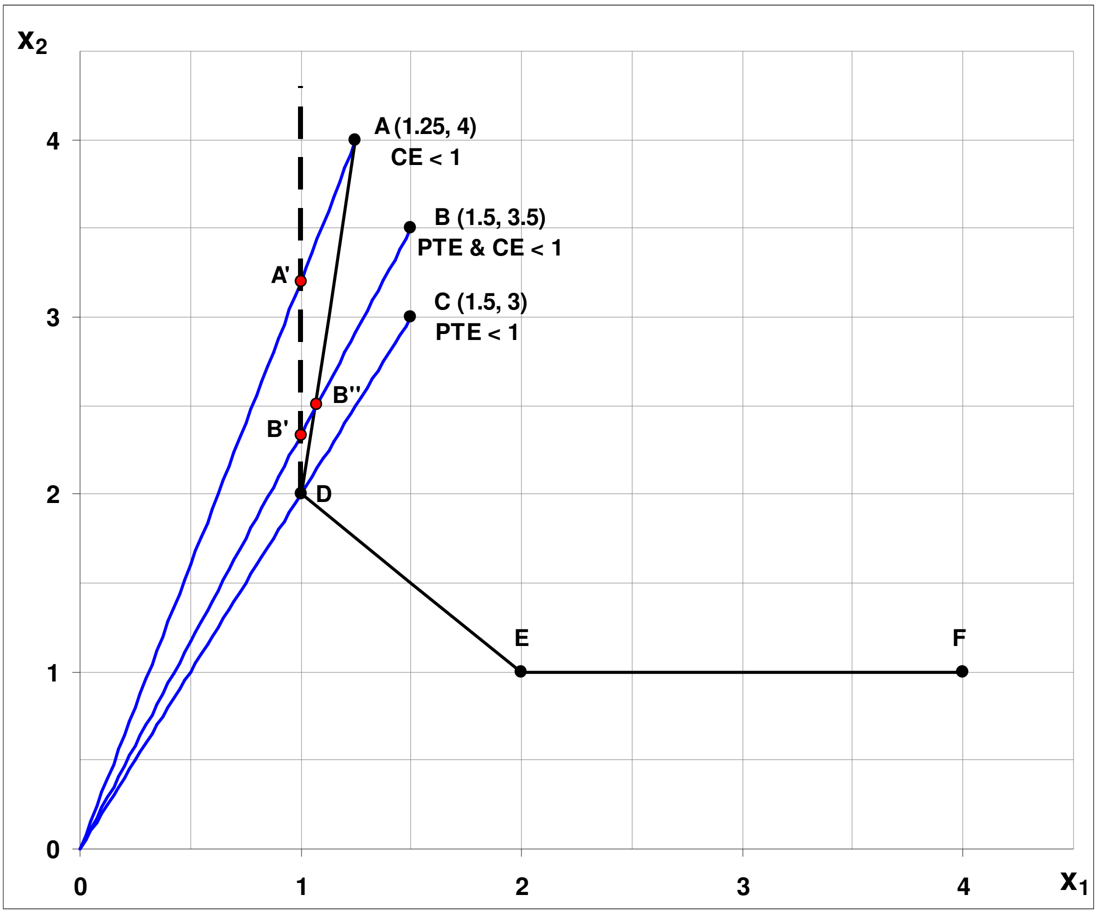

Fig. 1. Färe's approach (input-oriented, CRS)

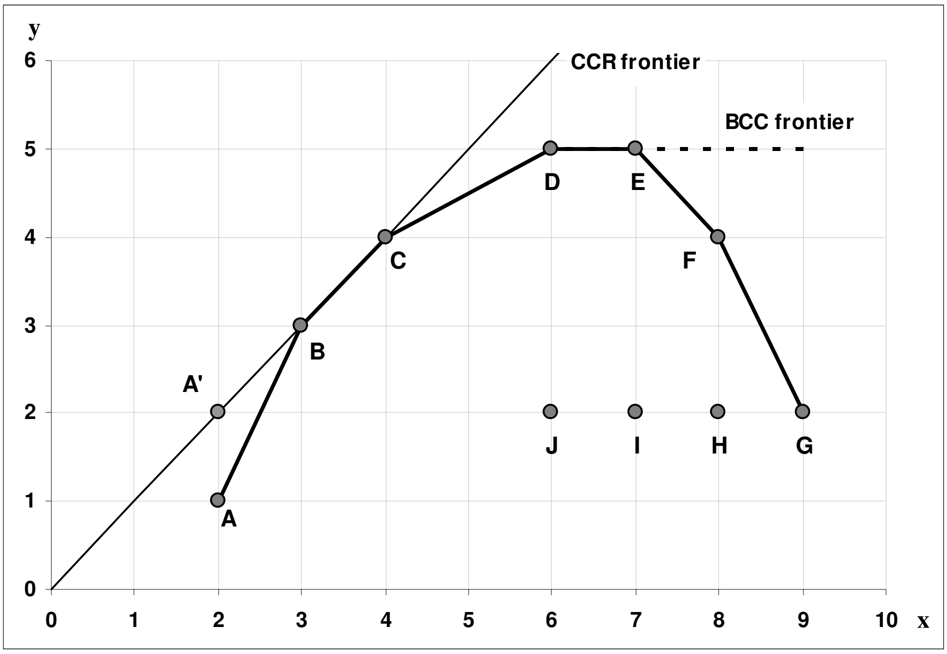

Fig. 2. DEA models and congestion

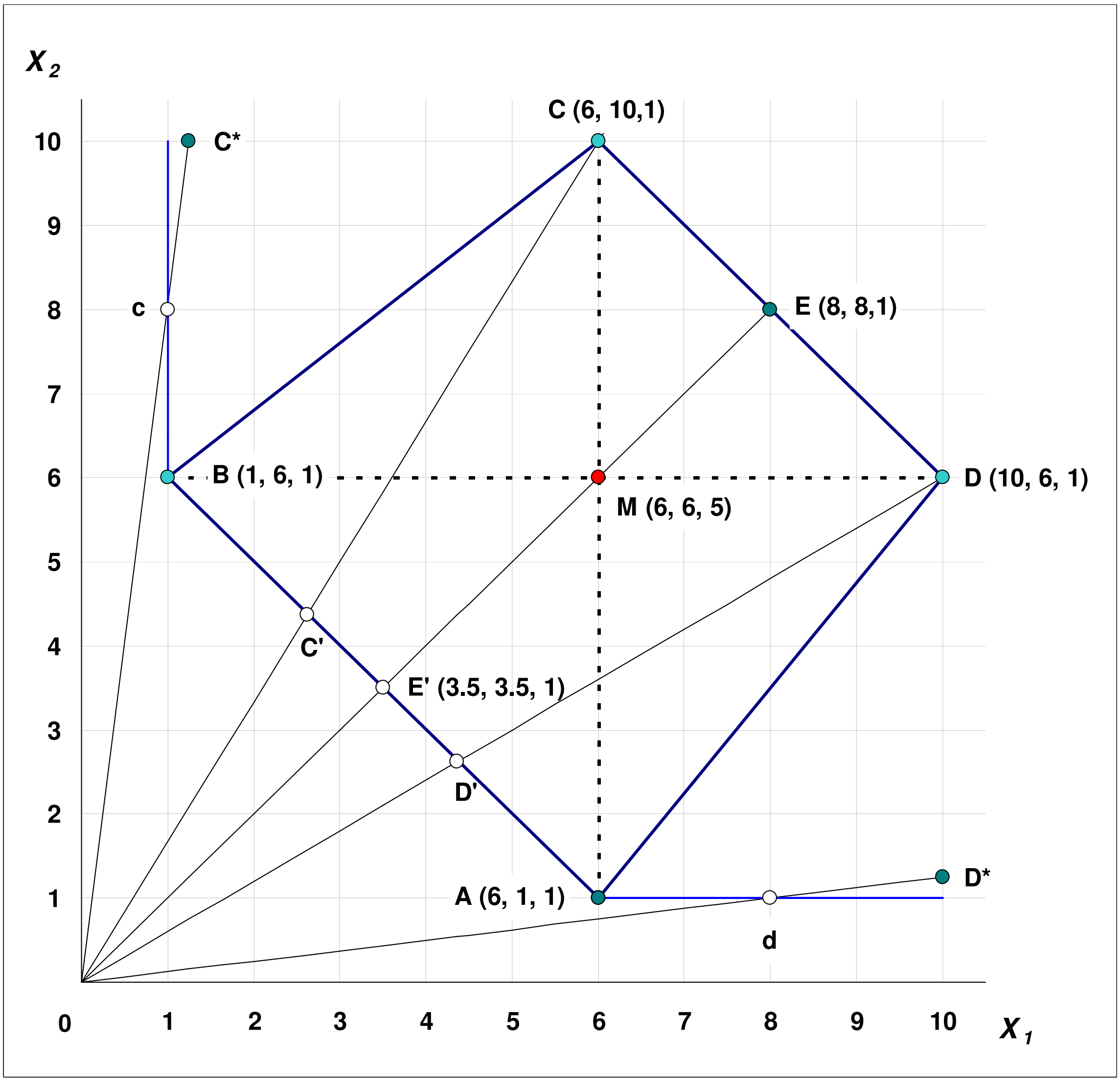

Fig. 3. An illustrative example

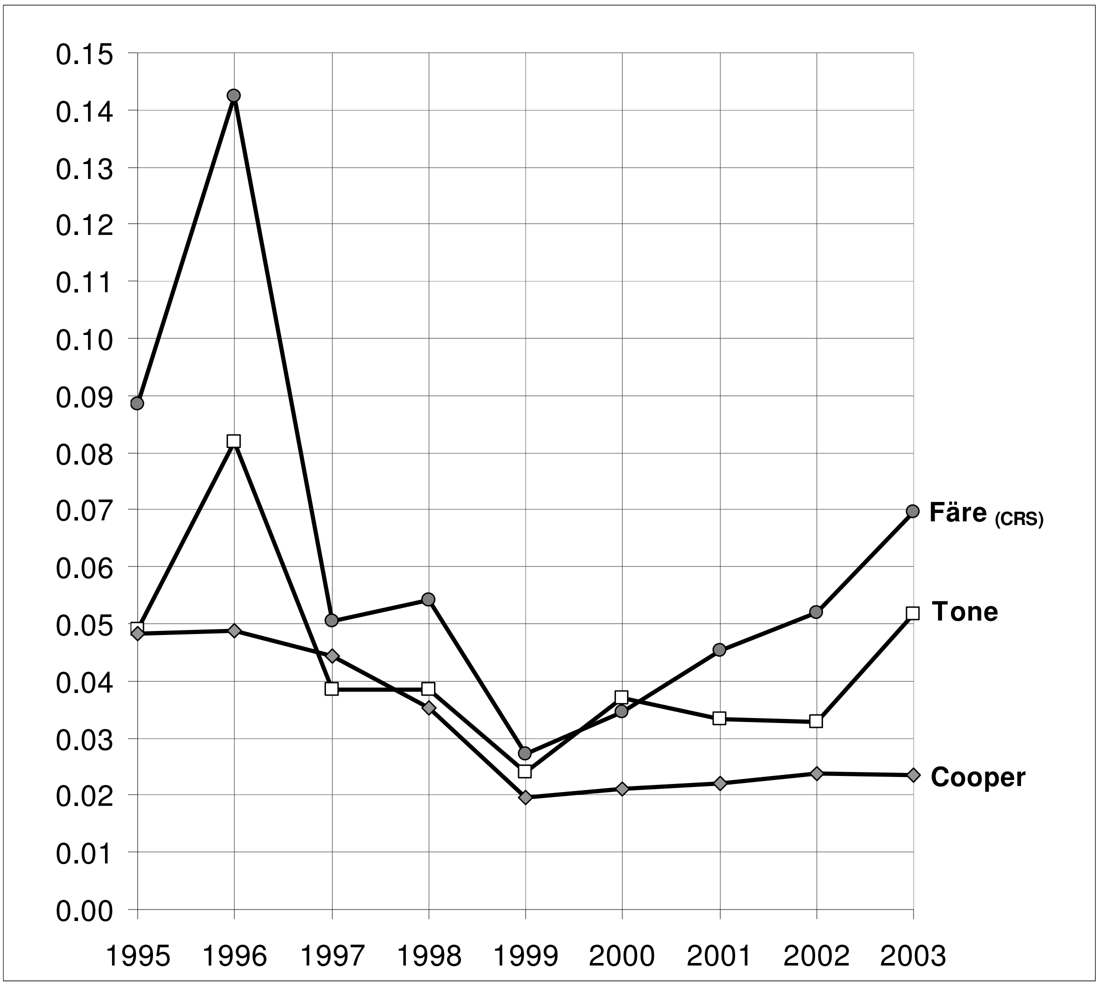

Fig. 4. Unweighted mean congestion scores (all universities)

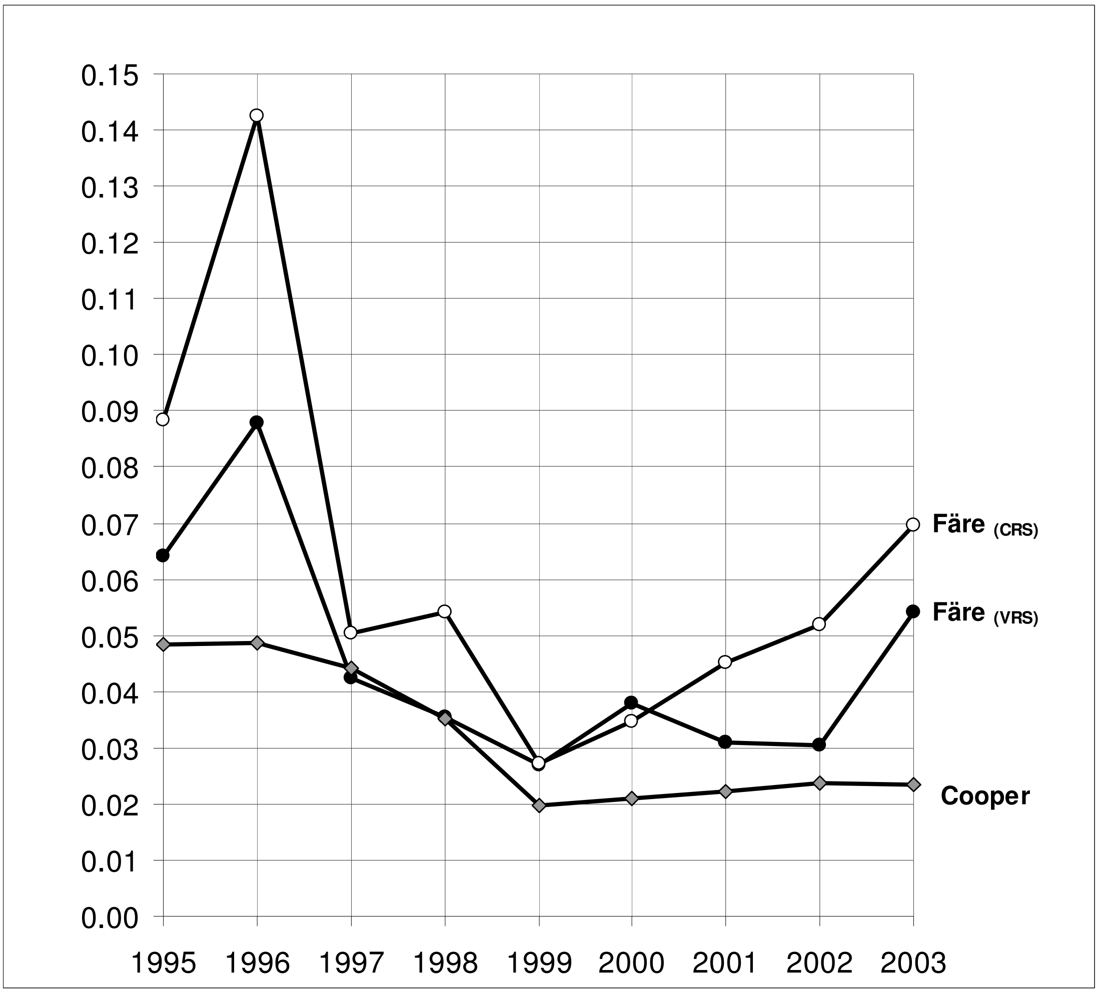

Fig. 5. Unweighted mean congestion scores (all universities)