Journal of Transport Economics and Policy, Volume 55, Part 4, October 2021, pp. 356–373

Bias in Transport Efficiency Estimates Caused by Misspecified DEA Models

Darold Barnum, Jason Coupet, John Gleason, Abagail McWilliams, and Annaleena Parhankangas

Address for correspondence: Jason Coupet, Department of Public Administration. School of Public and International Affairs, North Carolina State University, 209D Caldwell Hall, Raleigh NC 27695 USA (jacoupet@ncsu.edu).

We appreciate financial support provided by Dean Michael Mikhail, College of Business Administration, University of Illinois at Chicago, and the proprietary data on freight railroads by the Association of American Railroads. We thank Stanley Sclove and Georgios Karras for statistical/econometric assistance, and Finn Førsund for extensive help as we formulated our initial thoughts on the concepts involved. Extensive email correspondence regarding an *Applied Economics* article (Barnum *et al*., 2016a) between this paper's first author and Timo Kuosmanen caused us to do a great deal more work in justifying the use of our current CCR-NST and BCC-NST models. The referee offered very helpful suggestions that substantially increased our precision and clarity, as well as other matters. We alone are responsible for any misinterpretations of advice and advice not taken, any remaining lack of adequate justification, and all other shortcomings and deficiencies.

Abstract

One of the assumptions of Data Envelopment Analysis (DEA) in transportation studies is the substitutability/transformation of inputs/outputs. In this paper we examine three transport modes that frequently have employed DEA to measure technical efficiency. We ascertain whether input substitutions and output transformations are present in the transport data, and the impact on transport DEA scores when substitutions or transformations are not present. We then propose methods for correcting substitution and transformation specification errors. Future transport DEA studies should test for substitution and transformation in their data, to avoid biased DEA scores and spurious second-stage regression results.

Final version: December 2020

1.0 Introduction

Data Envelopment Analysis (DEA) is increasingly being used to measure efficiency in transport organisations. Cavaignac and Petiot (2017) report that in the 2008 to 2016 period, there were more than 37 transport DEA articles published annually; 2016 was the high point with 54 articles. The *Journal of Transport Economics and Policy* published DEA articles involving airlines (Yu *et al.*, 2019), airports (Gutiérrez and Lozano, 2016), rail (Oum and Yu, 1994; Growitch and Hetzel, 2009), and urban transit (Georgiadis *et al.*, 2014).

DEA uses linear programming to apply microeconomic production theory to sample data collected from actual organisations. In articles involving transport, authors are almost all transportation experts applying DEA, with only a few being DEA experts analysing transport (Cavaignac and Petiot, 2017). This distinction is important because DEA requires critical economic assumptions about the data in the sample being used. If the sample data do not comply with these economic specifications, it results in invalid efficiency scores**.**

Two essential economic specifications required by DEA are that, within the sample data, inputs must have been substituted for each other and outputs must have been transformed into each other. That is, a fixed set of outputs must have been produced with various efficient combinations of the inputs (represented by an isoquant in the two-input case); and a fixed set of inputs must have produced various efficient combinations of the outputs (represented by a production possibility curve in the two-output case) (Charnes *et al.*, 1978; Färe *et al.*, 1994; Cooper *et al.*, 2007; Fried *et al.*, 2008). The DEA axiom requiring this is the Convexity of the Production Set T Postulate[1] (Banker *et al.*, 1984, p. 1081):

$$\text{If } (\mathbf{X}_j, \mathbf{Y}_j) \in \mathrm{T}, j = 1, \ldots n, \text{ and}$$

$$\lambda_j \geqslant 0 \text{ are nonnegative scalars such that } \sum_{j=1}^{n} \lambda_j = 1, \text{ then } \left(\sum_{j=1}^{n} \lambda_j \mathbf{X_j}, \lambda_j \mathbf{Y_j} \right) \in \mathrm{T}.$$

For the actual data sample being analysed, substitution rates are estimated from each pair of efficient frontier inputs, and transformation rates are estimated from each pair of efficient frontier outputs. These estimated rates are converted into the weights used for the inputs and outputs of the Decision Making Unit (DMU) under consideration. Even if the required relationships exist in theory, they must also be present in the actual sample data for the weights and thereby efficiencies to be estimated validly.

There is sparse empirical evidence about whether substitutions and transformations are present in transport data sets, or about the impacts on efficiency scores when they are absent. Only one article presents such evidence, for a subgroup of one mode. Barnum and Gleason (2011) examined a data set of large transit systems, where substitutions and transformations were not present for those inputs and outputs frequently used in transit DEAs. They found that DEA technical efficiency scores were badly biased, with no consistency in the amount of bias.

[1] The three remaining postulates (minimum extrapolation, ray unboundedness, and inefficiency) are inapplicable.

This paper reports on DEA technical efficiency score bias in transportation when substitutions or transformations are missing. We examine three transportation modes that have been the subject of many DEA articles: (1) airlines; (2) urban transit; and (3) railroads. We find substantial differences between technical efficiency estimates produced by correct and incorrect specifications in each of the three transport modes examined, using the inputs and outputs common to many studies of these three modes. We fear that there are similar problems in many, if not most, applications of DEA to transport data.

This paper illustrates the specification errors using the two most prevalent DEA models; namely the Charnes-Cooper-Rhodes (CCR) constant-returns-to-scale model (Charnes *et al.*, 1978) and the Banker-Charnes-Cooper (BCC) variable-returns-to-scale model (Banker, 1984; Banker *et al.*, 1984). Substitutions and transformations are of course also necessary by more complex models that adapt CCR and BCC for specific purposes, as is the case for many recent transportation articles utilising DEA; so they also are subject to the same convexity postulate.

Finally, substitutions among inputs and transformations among outputs are also required by directional distance function and slack-based models. Both radial and non-radial formulations utilise the same production set, which they estimate from the sample data. But the radial and non-radial models would not be expected to yield the same efficiency scores since they estimate distances from the production set frontier differently (Tone, 2001; Färe and Grosskopf, 2004; Cooper *et al.*, 2011). That is, all will estimate invalid efficiency scores if substitution and transformation are not present in the sample involved, but their errors in efficiency estimates will almost certainly be different. Even Free Disposal Hull models require the presence of substitution and transformation, although of course not convexity (Tulkens, 1993).

This paper adds to the transportation body of knowledge about the presence and implications of misspecified DEA models concerning economic substitution and transformation requirements, and how to correct for their absence. We hope that future studies applying DEA to transportation data will identify and correct for specification errors.

2.0 Background

An initial motivation for DEA was to improve on efficiency models that simply compute the ratio of the summed normalised outputs to the summed normalised inputs for a Decision Making Unit (DMU). In such multicriteria models, each normalised input and output is equally weighted, unless unequal weights have been justified.

For an example of an equally weighted scheme, assuming three outputs and two inputs, the efficiency ratio for DMU_k would be:

$$Efficiency_k = (u_1y_1 + u_2y_2 + u_3y_3)/(v_1x_1 + v_2x_2).$$

Output weights are $u_1 = u_2 = u_3$ and input weights are $v_1 = v_2$, which means that the efficiency ratio would be the same whether or not the weights are used. No assumptions are necessary about the presence or absence of substitution among the inputs or transformation among the outputs. That is, even for DMUs with maximum efficiency in the sample at hand, increases in one variable may or may not necessitate decreases in another.

If decision makers want to assign differing values to inputs or outputs, they can replace the default weights of equal value with some scheme for representing the relative value of each input and output. Comprehensive taxonomies of such models can be found in Malczewski (1999), Malczewski and Jackson (2000), and Malczewski and Rinner (2015).

One key problem with letting all weights remain equal (as well as with other weighting schemes), as articulated by Charnes, Cooper, and Rhodes (Charnes *et al.*, 1978), is that they do not account for potential differences in the relative value of the various physical inputs and outputs for the *decision makers of the particular DMU involved.*

The insightful, economically sound solution that Charnes, Cooper, and Rhodes developed is to assume that physical inputs can be substituted for each other while maintaining constant output, and physical outputs can be transformed into each other while maintaining constant input. So weight differentials can be determined from the marginal inputs and outputs of the data sample for each individual DMU (Charnes *et al.*, 1978, p. 440).

If the data being used show that the efficient $\mathrm{DMU_k}$ is willing to give up five units of x_1 for one unit of x_2, then the Marginal Rate of Substitution (MRS) would be 5/1. So $\mathrm{DMU_k}$ values x_2 five times as much as x_1 and the weights reflect this valuation, $v_1 = 1/5$ and $v_2 = 1$. Plugging in these weights, the original denominator becomes $(0.2x_1 + 1.0x_2)$. The weights represent 'shadow prices' for a specific DMU, because they identify the relative value the DMU places on the physical inputs compared to each other (and the relative value of physical outputs compared to each other).

In order to estimate the different marginal rates of substitution for each DMU, we must use the data set that includes all DMUs to be analysed. As discussed in Section 3, we cannot employ DEA to determine whether substitution has occurred among the sample's inputs. However, once substitution has been proved to be present in our sample, then we can use DEA to estimate the MRS of the inputs of each DMU, and from these estimates determine the relative weights (*viz*., shadow prices) to assign to each input for the DMU in question. The relative input weights for each DMU depend solely on the MRS of the two inputs in question for that DMU.[2] Likewise, the relative output weights for each DMU depend solely on the Marginal Rate of Output Transformation (MRT) of the two outputs in question for that DMU. The same applies when there are more than two inputs and two outputs.

We think that much of the confusion about these concepts has grown out of the fact that it is often stated that, using the black box of linear programming, the DEA model whips through all of the possible weight combinations in order to find the set that yields the highest efficiency possible for the DMU in question. This of course is true, but that set *always* consists of weights that are the inverse of the MRSs and MRTs involved. So, in truth, linear programming is just the most efficient way of locating the applicable MRS and MRT values.

In two-input cases, for example, the ratio of input weights is equal to the inverse of the Marginal Rate of Input Substitution (MRIS) of the isoquant facet on which the input set

[2] Radial DEA models assume that the MRS of inputs for an inefficient DMU is the same as the MRS for an efficient DMU that uses the same proportions of each input, and the MRS of efficient DMUs are equal to the absolute value of the slope between the two inputs in question; and, of course, the relative weights of the two inputs will be the inverse of their MRS. The same is true for more than two inputs, and for outputs' MRT exchanges.

resides for DMUs on the efficiency frontier, and for inefficient DMUs that use input proportions equal to those of the applicable frontier facet. In two-output situations, the ratio of output weights is equal to the inverse of the (MRT) of the production frontier facet on which the output set resides for DMUs on the frontier, and for inefficient DMUs that use output proportions equal to those of that facet.

The input (output) frontiers are estimated from the piecewise isoquant (production possibility frontier) constructed from the sample being utilised. Therefore, DEA automatically replaces equal weights for each input, and for each output by differential weights for each individual DMU based on its own specific MRS and MRT — obtained from the applicable facet on the piecewise isoquant and production frontiers that have been estimated from the sample data being used.

As a result, conventional radial DEA models yield valid efficiency measures *only* when substitution among inputs and transformation among outputs are present in the sample being used. If there is no substitution (transformation), then no valid marginal rate of substitution (transformation) can be estimated, so there is no justification for replacing the original (equal) weights with differential weights. For a rigorous mathematical proof of this conclusion, which Charnes *et al.* (1978) illustrate on p. 440 of their seminal article, please read Section 3 of Førsund (2013).

3.0 Methods and Models

3.1 Deciding if inputs have been substituted and outputs transformed

DEA linear programmes do not identify which variables have been substituted or transformed and which have not. The DEA programmes simply identify (envelop) the extreme production units, based on whatever inputs and outputs happen to be used by the researcher.

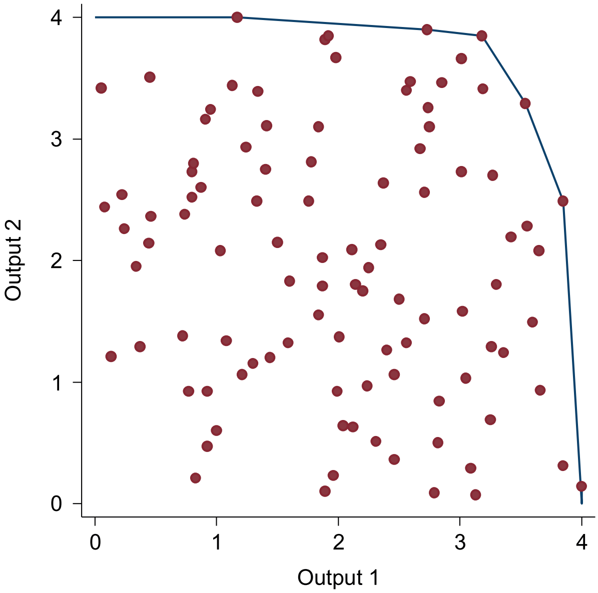
If two outputs (inputs) are unrelated variables, DEA still constructs archetypal efficiency frontiers, even though frontier increases in one output (input) have not been accompanied by decreases in another output (input). That is, even where there has been no transformation (substitution) and therefore no valid MRT (MRS) values. Consider a scenario where there is one input that is equal for all DMUs and two outputs that are completely unrelated to each other[3] (Figure 1).

As can be seen, DEA would construct a conventional piece-wise output transformation frontier, even though there are no transformations between the outputs of the DMUs on the false frontier. If we re-label Figure 1's axes as Input 1 and Input 2, then DEA constructs a classic (also false) piece-wise isoquant by connecting the extreme DMUs in the lower left.

Therefore, even in the absence of substitution (transformation), conventional radial DEA models will *always* construct a mathematical solution equivalent to a false frontier, and then proceed to compute the efficiencies of the DMUs using false slopes (that is, false MRSs and MRTs) to determine (invalid) input and output weights, and thereby invalid efficiency scores.

[3] They are independent random variables.

Figure 1
Independent Random Outputs and False Output Transformation Frontier



It is up to analysts to identify specification errors first, then adapt their DEA model to account for them, and only after that run the DEA. For example, if a conventional CCR programme assigns identical weights to the normalised inputs from a given DMU, this fact does not necessarily mean that they are perfect substitutes; they may or may not have been substituted for each other. Likewise, if the conventional DEA programme assigns different weights to normalised inputs, this does not imply that they are imperfect substitutes because, again, substitution may or may not have occurred. The meaning of the DEA weights can *only* be inferred after one has determined the presence (or absence) of substitution/transformation by means other than DEA. If there has been no substitution or transformation, then there is no economic justification for assigning unequal weights. Also, if the analyst manually assigns equal weights to two variables, this assignment does not imply perfect substitution/transformation, but only identifies the lack of any empirical evidence that substitution or transformation are present. Equal weights simply result in normalised variables' virtual values retaining the same relationships as their original values.

To determine any differences between the outcomes of misspecified and correctly specified DEA models, we must first perform statistical analyses to determine whether substitutions and transformations are present or absent for the specific sample involved. Then, where they are absent, we can compare conventional DEA results with those from models accounting for their absence.

3.2 Statistical tests for deciding if inputs have been substituted and outputs transformed

There is no need for statistical testing before choosing a correctly specified DEA model, when it is indisputable that efficient DMUs must consume inputs or produce outputs in fixed ratios — that is, a fixed proportion technology. Outputs of chicken legs and chicken wings (National Chicken Council, 2015) are an example of a fixed proportion output technology. Inputs of frames and wheels, where the output is bicycles, is an example of a fixed input proportion technology.

Statistical testing is necessary in samples where inputs (outputs) could be exchanged for each other but may not be in the sample at hand, or where the possibility of substitution or transformation is disputed. Parametric statistical methods offer the most powerful evidence that substitution/transformation has or has not occurred. Although there are many models that could be used to test for the presence of substitution or transformation, we use two parametric regression models developed elsewhere (Barnum *et al.*, 2016a; Barnum *et al.*, 2016b), with the input and output equation sets summarised next.

3.2.1 Inputs

One appropriate test of whether substitution exists among a set of inputs is to transform all inputs $x_1, \ldots, x_N$ and outputs $y_1, \ldots, y_M$ into their logarithms, and estimate the inputs' relationships by regressing each input on the other inputs and all outputs, with outputs serving as control variables (Barnum *et al.*, 2016a; Barnum *et al.*, 2016b). Suppose there are $j = 1, \ldots, \mathrm{J}$ observations, $n = 1, 2, \ldots, \mathrm{N}$ inputs, and $m = 1, 2, \ldots, \mathrm{M}$ outputs. Then Equation Set 1 would reflect the N required regressions. *Note that all but one input and all outputs appear on the RHS of each equation, and the one input absent from the RHS appears on the LHS. Each input appears on the LHS of one equation*:

$$\ln x_{1j} = \hat{\alpha} + \hat{\beta}_1 0 + \hat{\beta}_2 \ln x_{2j} + \hat{\beta}_3 \ln x_{3j} + \ldots + \hat{\beta}_N \ln x_{Nj} + \hat{\varphi}_1 \times \ln y_{1j} + \ldots + \hat{\varphi}_M \ln y_{Mj} + \varepsilon_j \quad \forall j, \tag{1.1}$$

$$\ln x_{2j} = \hat{\alpha} + \hat{\beta}_1 \ln x_{1j} + \hat{\beta}_2 0 + \hat{\beta}_3 \ln x_{3j} + \ldots + \hat{\beta}_N \ln x_{Nj} + \hat{\varphi}_1 \times \ln y_{1j} + \ldots + \hat{\varphi}_M \ln y_{Mj} + \varepsilon_j \quad \forall j, \tag{1.2}$$

$$\ln x_{Nj} = \hat{\alpha} + \hat{\beta}_1 \ln x_{1j} + \hat{\beta}_2 \ln x_{2j} + \hat{\beta}_3 \ln x_{3j} + \ldots + \hat{\beta}_N 0 + \hat{\varphi}_1 \times \ln y_{1j} + \ldots + \hat{\varphi}_M \ln y_{Mj} + \varepsilon_j \quad \forall j. \tag{1.N}$$

3.2.2 Outputs

If outputs are imperfectly transformed into each other, then they will reflect a statistically significant downward, curvilinear relationship, with the slope increasing as one moves right (east) from the origin. One procedure that can be useful to linearise this type of curvilinear relationship is to transform each output y_{mj} by computing the logarithm of a translation of its additive inverse as shown in equation (2) (when there are J observations and M outputs). The translation (for example, +2 in the equation) is required to ensure all numbers are greater than 1, so the log function can be applied. The additive inverse results in the log values approaching an upward linear relationship when a downward curvilinear relationship between the original variables exists. The log values do not produce a linear relationship if the additive inverse is not utilised (Barnum *et al.*, 2016a; Barnum *et al.*, 2016b). The transformed values of y, that is, y^T_{mj}, are:

$$y^T_{mj} = \ln(Max(y_{mj} \forall j \in \mathrm{J}) + 2 - y_{mj}) \ldots \forall j, m. \tag{2}$$

Then, regress each transformed output on the original values of the remaining outputs and all inputs, with inputs serving as control variables. Suppose there are $j = 1, \ldots, \mathrm{J}$ observations, $m = 1, 2, \ldots, \mathrm{M}$ outputs, and $n = 1, 2, \ldots, \mathrm{N}$ inputs. Then Equation Set 3 would reflect the M required regressions. *Note that all but one output and all inputs appear on the RHS of each equation, and the one output absent from the RHS appears on the LHS. Each output appears on the LHS of one equation*:

$$y_{1j}^{T} = \hat{\alpha} + \hat{\beta}_1 0 + \hat{\beta}_2 y_{2j} + \hat{\beta}_3 y_{3j} + \ldots + \hat{\beta}_M y_{Mj} + \hat{\varphi}_1 x_{1j} + \ldots + \hat{\varphi}_N x_{Nj} + \varepsilon_j \quad \forall j, \tag{3.1}$$

$$y_{2j}^{T} = \hat{\alpha} + \hat{\beta}_1 y_{1j} + \hat{\beta}_2 0 + \hat{\beta}_3 y_{3j} + \ldots + \hat{\beta}_M y_{Mj} + \hat{\varphi}_1 x_{1j} + \ldots + \hat{\varphi}_N x_{Nj} + \varepsilon_j \quad \forall j, \tag{3.2}$$

$$y_{Mj}^{T} = \hat{\alpha} + \hat{\beta}_1 y_{1j} + \hat{\beta}_2 y_{2j} + \hat{\beta}_3 y_{3j} + \ldots + \hat{\beta}_M 0 + \hat{\varphi}_1 x_{1j} + \ldots + \hat{\varphi}_N x_{Nj} + \varepsilon_j \quad \forall j. \tag{3.M}$$

Because we use an inverse of the left-hand output (y_{mj}^{T}), the relationship will be a positive one instead of a negative one for two outputs that are being transformed. The transformation will approximately linearise a curvilinear relationship concave from below, so the presence of output transformation between each set of two outputs will be identified by a statistically significant, positive linear relationship.

3.3 Transforming original variable values into common units of measure

In order to adapt radial DEA models such as the CCR and BCC models for cases lacking substitution or transformation, the involved sample variables must be transformed to common dimensionless units (that is, normalised) before they can be aggregated, which can easily be done with the use of ratios. It is conventional to normalise the units of variables by dividing each variable value by the sample mean of that variable (Washington *et al.*, 2010), so the common units are the ratio of a given value to its variable's mean (equations 4 and 5). We follow that convention, with transformed variables identified with the superscript τ:

$$x_{nj}^{\tau} = x_{nj} \Big/ \left(\sum_{j=1}^{J} x_{nj}/J \right) \quad n = 1, 2, \ldots, N;\ j = 1, 2, \ldots J, \tag{4}$$

$$y_{mj}^{\tau} = y_{mj} \Big/ \left(\sum_{j=1}^{J} y_{mj}/J \right) \quad m = 1, 2, \ldots, M;\ j = 1, 2, \ldots J. \tag{5}$$

3.4 DEA model used[4]

The conventional CCR and BCC models require that each DMU's input weights v_n and the output weights u_m be assigned based on the applicable frontier MRSs and MRTs, which must be used in order to maximise the DMU's technical efficiency scores, as discussed earlier. The multiplier form of the ubiquitous Banker-Charnes-Cooper (BCC) input-focused model is shown in linear programme (6.1)–(6.4). The programme maximises the DEA technical efficiency score for k', the specific organisation of interest; sometimes called the focus DMU or the target DMU. There are M outputs, N inputs, and J

[4] This model is more fully explained and justified in Barnum and Gleason (2011).

DMUs. The outputs y and inputs x are set exogenously, with endogenous input and output weights. The third endogenous variable z adjusts for variable returns to scale; this BCC model can be changed into the CCR model with the constraint $z = 0$, or simply by removing z from the model (Cook and Zhu, 2008; Cook and Zhu, 2015).

Adding to the basic BCC model, suppose there are two non-substituted inputs (in this example x_1 and x_2) and two non-transformed outputs (in this example y_1 and y_2). This means that there are no frontier values for the MRSs between the two inputs and MRTs between the two outputs, so there is no justification for assigning differential weights to normalised variables.

It is very important to remember that for a given sample, true substitution (transformation) will be assigned the same weights as the absence of substitution (transformation) by the conventional CCR and BCC models. Therefore, the presence of any given set of weights, whether assigned by operation of the conventional model or by additional constraints, yields no information about whether substitution (transformation) actually exists in the data set involved:

$$\max \quad \phi_{k'} = \sum_{m=1}^{M} u_m y^{\tau}_{mk'} + z_{k'} \tag{6.1}$$

$$\text{subject to} \quad \sum_{m=1}^{M} u_m y_{mj} - \sum_{n=1}^{N} v_n x_{nj} + z_{k'} \leqslant 0 \quad \forall j, \tag{6.2}$$

$$\sum_{n=1}^{N} v_n x^{\tau}_{nk'} = 1, \tag{6.3}$$

$$u_m, v_n, y^{\tau}_{mj}, x^{\tau}_{nj} > 0, \mathrm{z}_{k'} \text{ free} \quad \forall m, n, j, \tag{6.4}$$

$$v_1 - v_2 = 0, \tag{7.1}$$

$$u_1 - u_2 = 0. \tag{7.2}$$

Linear programme (6.1)–(6.4), (7.1)–(7.2) maximises the value of ϕ after assigning equal weights to inputs (outputs) that have not been substituted or transformed (7.1 and 7.2). Any inputs or outputs that have been substituted (transformed) would remain unconstrained. This programme can easily be adapted to any number of inputs and outputs, some of which replace each other and others which do not, with constraints added only for the non-replaced variables. When used instead of conventional CCR or BCC models, we designate the corrected models as CCR-NST and BCC-NST, with NST being an acronym for **N**o **S**ubstitution or **T**ransformation.

3.5 Alternative models

If *none* of the inputs have been substituted and/or *none* of the outputs transformed, there are alternative DEA models that can be used after all variables have been normalised. One is the Profit Efficiency Model (Cooper *et al.*, 2007, p. 260), with all input and output prices (weights) set to the same value, such as all being set to α. When all input and output price weights are the same, then the model measures technical efficiency of the physical inputs and outputs rather than economic efficiency of their monetary values.

If the normalised inputs have not been substituted but the outputs have been transformed, then one can use the Cost Efficiency model constraining only input prices to be equal (Cooper *et al.*, 2007, p. 258). If the normalised outputs have not been transformed but the inputs have been substituted, then the Revenue Efficiency model is appropriate, with output prices constrained to be equal (Cooper *et al.*, 2007, p. 260).

A non-DEA based model (Fixed Proportion Ratio model) that yields virtually identical results to the CCR-NST model, when there are no substitutions or transformations, may be examined in Barnum and Gleason (2011, Section 4.1). This model is valid for both fixed proportion technologies, and for cases where there are no isoquant and production frontiers.

Finally, a variety of other ways of dealing with the lack of substitution or transformation have been suggested, most of which can be applied without adapting the conventional models (Barnum and Gleason, 2011, Section 6.2).

We think that the relationships between the CCR and CCR-NST, and BCC and BCC-NST models, are clearer with constraint weighting terminology than would be the Cost, Revenue, and Profit Efficiency models, and the Fixed Proportion Ratio model. They tie the two models back to the role of the MRS and MRT in setting the weights in the DEA model, and the fact that when there is no substitution/transformation, there is no justification for assigning different weight ratios. Also, in the not uncommon case where some of the inputs (outputs) are substituted while others are not, it makes this clear by showing that weights between a set of inputs or outputs are constrained when substitution or transformation is absent, and not constrained when they are present. Thus, we have used them in the rest of this paper.

More generally, if there is no substitution and/or no transformation, then a generic DEA model is not the correct one to use. One must explicitly identify the nature of the production relations in order to determine the correct model[5] (Førsund, 2009). In addition, one must statistically test the data in the sample being used to justify the DEA model chosen.

4.0 Results: Application to Transportation Data Sets

For this paper, we assume that the articles' original inputs and outputs correctly reflect the production relationships, except for the potential absence of substitution and transformation. If we did not make this assumption, then it would be impossible to compare the specific impact of non-substitution and non-transformation on the DEA scores of the extant models. For example, if we argued that in addition to the lack of substitution and transformation a model was otherwise misspecified, and substituted what we felt was a correctly specified model, then we would be unable to identify the separate impacts of non-substitution or non-transformation on the original models being analysed.

Further, since many authors measure technical efficiency using monetary variables for inputs and outputs in place of physical units, we do the same where these instances occur.

[5] We are indebted to Finn Førsund for personally pointing this out to us.

Again, if we did not do so, then it would be impossible to compare the specific impact of non-substitution and non-transformation on the DEA scores of the extant models.[6]

4.1 Airlines

We used the inputs and outputs employed in a recent *TR-A* article on the topic of airline efficiency (Barros *et al*., 2013), with similar variables being common in many other airline DEAs. Inputs are: (1) total costs; (2) number of FTE employees; and (3) gallons of fuel. The three outputs are: (1) total revenue; (2) revenue passenger miles; and (3) passenger load factor.

Except for using Barros *et al*.'s choice of inputs and outputs, our analysis departs from most other characteristics of that article. We use a different sample, and do not employ the B-convex DEA model that was the focus of that paper. However, since the authors used a Variable Returns to Scale (VRS) model, to compare it to their results we use the VRS BCC model. The source of our data is the MIT Airline Data Project (MIT Global Airline Industry Program, 2015), which we discovered thanks to the Barros *et al*. article. In order to maximise our sample size to yield the best chance of identifying substitutions and transformations, we used all of the data points available from 1995 until 2014 that had a complete set of the inputs and outputs. This resulted in a sample of 251 observations for the DEA computations.

4.1.1 Statistical analysis of input and output relationships

We tested the relationship of each pair of inputs, controlling for the third input and the three outputs (equation 1). There was no substitution among the inputs, and, using equation (3), the only statistically significant transformation was between Revenue and Load Factor. When transformed Revenue was the response variable, $t = 3.38$, $P > |t| = 0.001$. So the weights of the standardised inputs must be held equal in the DEA, as must the weights of the standardised outputs (except that the fact that there was transformation between Revenue and Load Factor means that we do not restrict their weights relative to each other).

As discussed earlier, if there is a consensus that certain inputs cannot be substitutes or that certain outputs cannot be transformed, then there is no reason to carry out statistical testing and the researcher can simply treat their non-substitution or non-transformation as given. In our current sample, we think it highly unlikely that management could decrease employees and replace them by increasing fuel, or vice versa. In addition, there is overlap in the cost input with the other two inputs, even though they are being measured with different units. Thus, our statistical results support the scenario that is virtually certain to be true — that there can be no substitution among the inputs. Thus, whether or not one wishes to argue with our statistical estimates, it must be concluded that substitution could not be present and therefore a DEA model adjusting for non-substitution must be employed.

[6] Monetary units measure total efficiency rather than technical efficiency, since they incorporate technical, allocative, factor price, and product price efficiencies. However, monetary units measure technical efficiency if we assume that factor prices per unit are equal across all inputs and all DMUs, and that product prices per unit are equal across all outputs and all DMUs. We infer that this assumption has been made by authors using monetary values for inputs or outputs, which only requires making the ubiquitous assumption that everything other than the variables under examination are equal.

The outputs are trickier because management could increase one output while decreasing another by adjusting prices, so it would seem prudent to conduct appropriate statistical tests to see if this has occurred in our sample. This might explain the puzzling (to us) transformations between Revenue and Load Factor; increasing prices might possibly, in some circumstances, increase Revenue and decrease Load Factor. Therefore, if it were possible (although unlikely) for substitution or transformation to occur, then we counsel to do the statistical testing.

4.1.2 BCC and BCC-NST scores

When comparing the standard BCC scores of the model containing no weight restrictions and the BCC-NST model that corrects for the absences of substitutions and transformations, the R-squared statistic was only 0.235, meaning the conventional BCC efficiency scores are very bad predictors of the DMU's true technical efficiencies. Moreover, there are sharp changes in the ranks of many of the 251 DMUs. Seven DMUs had rank increases of over 100, and 32 DMUs had rank decreases exceeding 100. In addition, while the conventional BCC model reports that 39 DMUs were efficient, in truth only 11 were.

We cannot say whether these findings would apply with different inputs and outputs, but they ought to raise concerns that appropriate statistical tests of input substitution and output transformation should be conducted before proceeding with an airline DEA analysis.

4.2 Transit

A 2011 analysis used the 66 largest US bus transit systems (excluding the New York MTA) as the DMUs, and six years of data for each system, with contemporaneous DEAs computed for each year. As has been typical for many transit DEA articles, inputs were fuel, labour, and buses, and outputs were vehicle miles and passenger miles, which in truth represent a classic case of a fixed proportion technology.[7] The complete results may be found in Barnum and Gleason (2011). Briefly, they reported that conventional DEA yielded badly biased efficiency estimates. The R-square between the conventional and correct DEA models averaged only 58 per cent.

For this study, we analyse the efficiency of 46 bus routes of a US transit agency, treating each route as a DMU. We obtained the data from the original authors (Barnum *et al.*, 2008). The article employs a 'reverse' two-stage methodology, under which the external influences on the inputs and outputs are removed from the data first, then the DEA is conducted with data already adjusted for the external influences, using a protocol developed by Barnum and Gleason (2008).

[7] As is typical in transit DEAs (which also is the case for other DEAs), factors and products not included in the analysis are assumed to be equal across DMUs. So if the only outputs included are vehicle miles and passenger miles, this does not allow for other possible inputs or outputs that might change the passenger-miles-to-vehicle-miles ratio. If one wishes to allow fare or load factor or bus size to vary, for example, then they would have to be included as outputs or otherwise accounted for. Further, for an efficient property, DEA theory requires that increases in vehicle miles lead to decreases in passenger miles, which is illogical. Empirical evidence shows the vehicle and passenger miles increase or decrease together, everything else equal, and therefore represent a fixed product proportion technology.

Inputs	Seat hours Seat kilometres
Outputs	Number of unlinked passenger trips Daily time span of service Average frequency of service Maximum frequency of service On-time performance percentage Population
Second-stage regression variables	Population density Population*key route Title 6 route
Second-stage regression model	OLS regression, with robust variance-covariance matrix

4.2.1 Statistical analysis of input and output relationships

In the original article, the authors did not check to see whether the two inputs act as substitutes, or if some or all outputs have been transformed. We do so here; as usual, the data were normalised using equations (4) and (5), and we use the functional forms identified in equation (1) for inputs and equation (3) for outputs. We regressed the log of seat kilometres on the log of seat hours, holding the logs of the five outputs constant as control variables. Likewise, we compared the various pairs of outputs, holding the remaining outputs and both inputs constant.

There was no substitution between seat kilometres and seat hours, with a regression coefficient of +1.18 and a t-statistic of 15.74 (for input substitution, the coefficient must be *negative* to a statistically significant degree). For outputs that have been transformed, recall that the regression coefficient connecting the output serving as the response variable to any of the outputs serving as independent variables (treating other outputs and all inputs as control variables) must be positive to a statistically significant degree (since the response output is the additive inverse of that output's original value). In only one case did this occur; when the 'Average Frequency of Service' was serving as the response variable, the regression coefficient of the 'Daily Time Span of Service' was +0.16 with $t = 3.38$, $\mathrm{P} > |t| = 0.0004$. The presence of such substitution is exactly what would be expected, so the relationship is not surprising. Therefore, the DEA weights of the two normalised inputs must be held equal, as must the DEA weights of all pairs of outputs except between the outputs Average Frequency and Time Span.

As with airlines, it would appear certain that the inputs and most of the outputs could not be substituted/transformed for each other for technological or logical reasons. Therefore, whether or not one agrees with our statistical methodology, it is clear that valid DEA scores can only be estimated with a model that corrects for non-substitution and non-transformation.

4.2.2 CCR and CCR-NST scores

When comparing the CCR scores of the model containing no weight restrictions and the CCR-NST model containing all appropriate weight restrictions, the R-squared statistic was only 0.451, meaning the CCR efficiency scores were a very bad predictor of the DMU's true technical efficiencies. Perhaps worse, 11 of the 46 routes were reported efficient by the unconstrained DEA, when in truth only one of their number was actually efficient. All but one of the 11 routes were frequently used as part of a 'best practice' set with which inefficient routes were compared, which means that virtually all recommendations for efficiency improvement were flawed.

4.2.3 Second-stage regression

Although the original research removed the influence of exogenous factors from the inputs and outputs before conducting their DEAs, we used the original inputs and outputs for our DEAs, so we could apply the normal two-stage procedure in which the DEA scores are regressed on independent variables to measure the influence of the environment on the scores. In our regressions, we used the same exogenous variables as the original research. We compared the results when the unconstrained CCR scores were used as the response variable with the results, when the constrained CCR-NST scores were used as the response variable (Table 1).

As can be seen, the R-squared values do differ both substantially and significantly, and a Chow/Wald test reports that the two regressions differ to a statistically significant degree. Therefore, once again, not only are there substantial differences in the DEA efficiency scores between the CCR-NST and CCR estimates, but there are also both substantial and statistically significant differences between the second-stage regression estimates.

Table 1

Comparison of Unconstrained and Constrained DEA Models for Bus Route Efficiency

	DEA Score – CCR	*DEA Score – CCR-NST*	*Change*
Constant	0.1834 (0.656)	0.2826 (0.318)	0.0992 (0.842)
Population	−1.1463 (0.000)*	−0.8325 (0.000)*	0.3138 (0.255)
Population Density	1.1284 (0.016)	0.5190 (0.105)	−0.6094 (0.270)
Pop. * Key Route	0.5547 (0.001)*	0.0471 (0.659)	−0.5076 (0.006)*
Title 6 Route	−0.1398 (0.062)	−0.2431 (0.000)*	−0.1033 (0.252)
	R-squared = 0.446	R-squared = 0.626	

Note: Data (in parentheses) reports two-tail probability that the estimate above it differs from zero by chance; that is, the alpha error. For clarity, we have starred* those estimates statistically significant with a probability less than 0.01. The joint test of the intercept and regression coefficients of the two equations reports that they differ with a statistical significance of 0.000 according to the Chow/Wald test.

Source: Stata 13 OLS regression model, with robust standard errors and post-estimation use of the 'test' command. The constrained DEA model requires inputs to have equal weights and all outputs to have equal weights, except operating hours per day vs. average frequency.

4.3 Railroads

Lim and Lovell (2009) developed an intertemporal, decomposition model to identify the influences on short-run profit changes of US Class I (Freight) Railroads, between 1996 and 2003. The DEA efficiency scores, which were only a small part of the overall analysis, were estimated with linear programmes equivalent to the BCC model using sequential DEA. This results in efficiency measures for each railroad that include: (1) its efficiency in each sample year t compared to data for all railroads for the current sample year t and all prior years' data; and (2) its efficiency in each sample year t compared to data for all railroads for year $t+1$ and all prior years' data. For all railroads for all periods, this results in 125 measures of efficiency, all computed assuming the existence of substitution among the inputs. The sole output is Freight Revenue Ton-Miles. The three inputs are: (1) gallons of diesel fuel; (2) number of employees; and (3) variable maintenance expenses. Lim and Lovell excluded the quasi-fixed inputs of 'Way and Structure' and 'Equipment', arguing (correctly in our opinion) that they cannot be changed in the short run, so they have no effect on changes in short-run output. We collected our data for the railroads and time periods used by the original authors (Association of American Railroads data), which the AAR generously provided to us, but, although we used the same inputs and output, we have no way of knowing how closely our data values actually match Lim and Lovell's.

4.3.1 Statistical analysis of inputs and outputs

Using equation (1), none of the three inputs has been substituted for each other, so it would be inappropriate to use BCC with unconstrained input weights. There was only one output, so no analysis of output is necessary. Of course, our statistical findings merely confirm the obvious fact that fuel, employees, and maintenance could not be substitutes, so, regardless of the statistical findings, it is clear that a DEA model correcting for non-substitution must be used.

4.3.2 BCC and BCC-NST scores

The 125 DEA scores ranged from 0.68 to 1 when applying the BCC model to the data, and from 0.47 to 1 when applying the BCC-NST model; 40 of the 125 scores were reported as efficient by BCC, whereas only 21 of the original 40 remained efficient under the BCC-NST adaptation (the remaining 19 falsely efficient scores had true efficiencies ranging from 0.73 to 0.99). Pearson's R-square value between the two scores was 0.57. In summary, the (incorrect) BCC scores estimate only 57 per cent of the (correct) BCC-NST scores, and 19 of the 40 observations that BCC reported as efficient are, in truth, not efficient.

5.0 Discussion and Conclusions

A frequently mentioned benefit of Data Envelopment Analysis (DEA) is that it is based on very few assumptions (Banker, 1984; Charnes *et al.*, 1985). Nevertheless, to estimate valid technical efficiency scores, those few assumptions *must* be satisfied by the data in the sample being analysed. Most importantly, the convexity assumption requires that there have been substitutions between each pair of inputs and transformations between each pair of outputs in the sample data (Banker *et al.*, 1984).

Input substitution and output transformation are critical because the differential weights for inputs in the radial DEA linear programmes are determined solely by the Marginal Rates of Substitution (MRS) among the sample's inputs. Likewise, differential weights for outputs in radial DEA linear programmes are determined solely by the Marginal Rates of Transformation (MRT) among the sample's outputs. For a rigorous mathematical proof that the ratio of DEA input (output) weights is the inverse of the corresponding MRS (MRT), which Charnes *et al.* illustrate on p. 440 of their 1978 article, please read Section 3 of Førsund (2013).

MRS and MRT values are the sole determinates of differential input and output weights in radial DEA models. The weights represent 'shadow prices' for a given DMU's decision makers, because they identify the relative values of inputs compared to each other, and the relative values of outputs compared to each other. There is no economic justification for allowing differential weights to be assigned to variables that have not been interchanged.

In short, the data set must include substitutions among its inputs before differential weights can be validly estimated when the DEA aggregates virtual inputs. In addition, it must include transformations among its outputs before differential weights can be validly estimated when the DEA aggregates virtual outputs. Whether or not substitutions and transformations could occur for the population as a whole, if they do not occur in the sample being analysed then it is impossible to estimate valid MRS and MRT values that apply to that sample, thereby making it impossible to construct valid weights to be used by DEA.

DEA programmes do not identify which variables have been substituted for each other and which have not. They simply identify those production units in the sample data that are the most extreme and therefore envelop the remaining units in the sample. So, even if the inputs and outputs are unrelated random variables, DEA still constructs a (false) frontier and uses it to identify (false) MRS and MRT values, and then uses these false values to determine (invalid) input and output weights. Therefore, it is necessary to perform statistical tests on the raw input and output data to assure proper relationships exist before implementing DEA models.

In our statistical analysis of airline, transit, and freight railroad data samples, input substitutions and output transformations are absent between most of the variables. The absence of the required substitutions and transformations results in grossly misleading CCR and BCC estimates of technical efficiency in all three of our sample data sets, and affects second-stage regression results significantly.

We looked at three transport modes, using input and output variables ubiquitous to DEA articles involving their industries. All three of them resulted in major concerns about the substantial differences between conventional efficiency estimates and those yielded by the correct models. We believe that the same problems would undoubtedly be present in complex DEA models that are partly based on CCR or BCC, as well as in non-radial DEA models, although we do not speculate about the size of the impacts.

In addition to DEA studies of airlines, transit, and freight rail, we would expect the same pitfalls to be present in DEA studies of other modes of transport, including those more heavily involving fixed assets such as airports and seaports. These findings should raise concerns about any future transportation research that uses DEA if it does not first empirically verify substitutions among its inputs and transformations among its outputs, and that fails to adapt its DEA models when substitutions and transformations cannot be shown to be present.

References

Banker, R. D. (1984): 'Estimating the most productive scale size using data envelopment analysis', *European Journal of Operational Research*, 17, 35–44.

Banker, R. D., A. Charnes, and W. W. Cooper (1984): 'Some models for estimating technical and scale inefficiencies in data envelopment analysis', *Management Science*, 30, 1078–92.

Barnum, D. and J. Gleason (2011): 'Measuring efficiency under fixed proportion technologies', *Journal of Productivity Analysis*, 35, 243–62.

Barnum, D., J. Coupet, J. M. Gleason, A. McWilliams, and A. Parhankagas (2016a): 'Impact of input substitution and output transformation on data envelopment analysis decisions', *Applied Economics*.

Barnum, D. T. and J. M. Gleason (2008): 'Bias and precision in the DEA two-stage method', *Applied Economics*, 40, 2305–11.

Barnum, D. T., M. Johnson, and J. M. Gleason (2016b): 'Importance of statistical evidence in estimating valid DEA scores', *Journal of Medical Systems*, 40, 40–7.

Barnum, D. T., S. Tandon, and S. McNeil (2008): 'Comparing the performance of bus routes after adjusting for the environment, using data envelopment analysis', *Journal of Transportation Engineering*, 134, 77–85.

Barros, C. P., Q. B. Liang, and N. Peypoch (2013): 'The technical efficiency of US Airlines', *Transportation Research Part A: Policy and Practice*, 50, 139–48.

Cavaignac, L. and R. Petiot (2017): 'A quarter century of data envelopment analysis applied to the transport sector: a bibliometric analysis', *Socio-Economic Planning Sciences*, 57, 84–96.

Charnes, A., W. W. Cooper, and E. Rhodes (1978): 'Measuring the efficiency of decision making units', *European Journal of Operational Research*, 2, 429–44.

Charnes, A., W. W. Cooper, B. Golany, L. Seiford, and J. Stutz (1985): 'Foundations of data envelopment analysis for Pareto-Koopmans efficient empirical production functions', *Journal of Econometrics*, 30, 91–107.

Cook, W. D. and J. Zhu (2008): *Data Envelopment Analysis: Modeling Operational Processes and Measuring Productivity*, Delaware, CreateSpace Independent Publishing Platform.

Cook, W. D. and J. Zhu (2015): *Data Envelopment Analysis: Balanced Benchmarking*, Amazon Digital Services, Inc., Seattle, WA.

Cooper, W. W., L. M. Seiford, and K. Tone (2007): *Data Envelopment Analysis: A Comprehensive Text with Models, Applications, References and DEA-solver Software*, New York, Springer.

Cooper, W. W., L. M. Seiford, and J. Zhu (eds) (2011): *Handbook on Data Envelopment Analysis*, New York: Springer.

Färe, R. and S. Grosskopf (2004): *New Directions: Efficiency and Productivity*, Boston, Kluwer Academic Publishers.

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

Førsund, F. R. (2009): 'Good modelling of bad outputs: pollution and multiple-output production', *International Review of Environmental and Resource Economics*, 3, 1–38.

Førsund, F. R. (2013): 'Weight restrictions in DEA: misplaced emphasis?', *Journal of Productivity Analysis*, 40, 271–83.

Fried, H. O., C. A. K. Lovell, and S. S. Schmidt (2008): *The Measurement of Productive Efficiency and Productivity Growth*, Oxford, Oxford University Press.

Georgiadis, G., I. Politis, and P. Papaioannou (2014): 'Measuring and improving the efficiency and effectiveness of bus public transport systems', *Research in Transportation Economics*, 48, 84–91.

Growitch, C. and H. Wetzel (2009): 'Testing for economies of scope in European Railways. An efficiency analysis', *Journal of Transport Economics and Policy*, 43, Part 1, 1–24.

Gutiérrez, E. and S. Lozano (2016): 'Efficiency assessment and output maximization possibilities of European small and medium sized airports', *Research in Transportation Economics*, 56, 3–14.

Lim, S. H. and C. A. K. Lovell (2009): ‘Profit and productivity of US Class I railroads’, *Managerial and Decision Economics*, 30, 423–42.

Malczewski, J. (1999): *GIS and Multicriteria Decision Analysis*, New York, NY, Wiley.

Malczewski, J. and M. Jackson (2000): ‘Multicriteria spatial allocation of educational resources: an overview’, *Socio-Economic Planning Sciences*, 34, 219–35.

Malczewski, J. and C. Rinner (2015): *Multicriteria Decision Analysis in Geographic Information Science*, Berlin Heidelberg, Springer.

MIT Global Airline Industry Program (2015): Airline Data Project.

National Chicken Council (2015): ‘Wings 2 rule Superbowl XLIX’, *Meat & Poultry* (online). Available at www.meatpoultry.com/articles/news_home/Trends/2015/01/Wings_to_rule_Superbowl_XLIX.aspx?ID = {233AB6E4-5575-4666-9755-5D9255F689B1}&cck = 1 (accessed 9 July 2015).

Oum, T. H., W. G. Waters, and C. Yu (1999): ‘A survey of productivity and efficiency measurement in rail transport’, *Journal of Transport Economics and Policy*, 9–42.

Tone, K. (2001): ‘A slacks-based measure of efficiency in data envelopment analysis’, *European Journal of Operational Research*, 130, 498–509.

Tulkens, H. (1993): ‘On FDH efficiency analysis: some methodological issues and applications to retail banking, courts, and urban transit’, *Journal of Productivity Analysis*, 4, 183–210.

Washington, S., M. G. Karlaftis, and F. L. Mannering (2010): *Statistical and Econometric Methods for Transportation Data Analysis*, Boca Raton, Fla, Chapman & Hall/CRC.

Yu, H., Y. Zhang, A. Zhang, K. Wang, and Q. Cui (2019): ‘A comparative study of airline efficiency in China and India: a dynamic network DEA approach’, *Research in Transportation Economics*, 100746.