University of Connecticut

Department of Economics Working Paper Series

The Validity of Input Aggregation in DEA Models: A Statistical Test

Subhash C. Ray
University of Connecticut

Kankana Mukherjee
Worcester Polytechnic Institute

Working Paper 2005-54R

October 2005, revised November 2006

341 Mansfield Road, Unit 1063
Storrs, CT 06269–1063
Phone: (860) 486–3022
Fax: (860) 486–4463
http://www.econ.uconn.edu/

This working paper is indexed on RePEc, http://repec.org/

Abstract

We address the problem of input aggregation in DEA models within a broader framework and provide a direct link between input aggregation in DEA on the one hand and the test of parameter restrictions implied by input aggregation in an explicitly specified production function on the other. We show that when input prices vary across firms, the DEA LP problems for measuring efficiency scores of individual firms from the aggregated model have to be appropriately modified. An empirical application of the revised model using data from Indian manufacturing sector reveals that the validity of aggregating production and non-production workers into a composite labor input using Banker's F-test cannot be rejected.

Journal of Economic Literature Classification: C61, D21, L6

Keywords: Banker's F-test, data envelopment analysis, input aggregation, parameter restrictions

The paper has benefited from comments received from participants at the 2006 INFORMS Annual Meeting.

THE VALIDITY OF INPUT AGGREGATION IN DEA MODELS: A STATISTICAL TEST

Selecting the right number of distinct inputs and outputs for the specification of the underlying technology is often quite difficult in applied production analysis. In many cases, for a variety of reasons, several inputs (or outputs) may be aggregated into a single composite input (or output). In agricultural production, for example, the number of even the important crops may be too large to treat them as separate outputs. This requires aggregating them into a single index of crop output. Similarly, especially in developing countries, one constructs a measure of total cultivated area by adding up irrigated and unirrigated acreage, assigning a higher weight to irrigated land. Because there are farms that cultivate only unirrigated land, including irrigated land as a distinct input would create the usual problems associated with zero input levels. In manufacturing, quantities of different kinds of fossil fuels are sometimes aggregated in terms of their British Thermal Units (BTU) equivalent. In many cases, information about input use is available only in value terms. One may, for example, have to use the total wage payments instead of the numbers of different categories of employees to measure the labor input. This is especially true for applications where the input-output data are indirectly constructed from financial statements of firms.

In applications of Data Envelopment Analysis (DEA), maintaining a large number of inputs (or outputs) in the model may result in all or most of the decision making units (DMUs) to be rated as efficient. As explained by Leibenstein and Maital (1992) this is a result of the dimensionality of the input /output space relative to the number of observations. Meaningful aggregation of multiple inputs or outputs, therefore, reduces the number of constraints in the linear programming model, thereby increasing the "degrees of freedom". However, while input aggregation has some advantages when it is valid, it does pose problems in cases when it is not valid.

The theoretical implications of such aggregation have been addressed in some previous studies. For instance, based on Monte Carlo simulations, Thomas and Tauer (1994) argue that with linear aggregation of inputs the technical efficiency measure becomes an economic measure of efficiency comprising of both technical and allocative components. As such, this introduces bias in the measurement of technical efficiency. In a subsequent study by Tauer (2001), the results of simulations indicate that technical efficiency estimates computed by DEA are biased even when the exact aggregator function is used to aggregate inputs. Färe and Zelenyuk (2002) and Färe et al (2004) demonstrate that if some inputs are introduced in the DEA model in price

aggregated form, the technical efficiency measure will be *biased downward* as compared to if the inputs were included in disaggregated terms. Further, the bias is equal to the allocative inefficiency. They further illustrate that the DEA technique will yield unbiased technical efficiency scores even when inputs are price aggregated, if and only if there is no allocative inefficiency in the subvector of inputs that is aggregated.

The above theoretical developments reveal that while including too many distinct inputs (or outputs) in the DEA model may lead to problems associated with overspecification, aggregating inputs into a fewer number of inputs to be included in the model may lead to biased measures of technical efficiency. The validity of input aggregation in any specific application is therefore an empirical question. However, the statistical side of this issue has been far less explored. Of the few studies that address this aspect of the problem, some are nonparametric while others assume some parametric form of the statistical distribution. Simar and Wilson (2001) discuss bootstrap procedures for testing several restrictions in nonparametric models, including tests for input or output aggregation. Sirvent et al (2005) use simulations to evaluate the performance of a number of statistical tests that can be used for the selection of variables in DEA efficiency models. In a recent paper Banker et al. (2005) proposes a method of estimating allocative inefficiency when only aggregate cost or revenue data and quantity data (but not price data for individual inputs or outputs) are available. They provide a test of *allocative efficiency* through an application of Banker's (1993) F-test based on the DEA *technical efficiency* residuals from the aggregate and the disaggregated models.

In an empirical application, even when price and quantity data are all available, one might prefer to aggregate subgroups of inputs (or outputs) to overcome the 'curse of dimensionality'. There are, of course, various ways in which inputs may be aggregated. In some contexts the aggregation weights may be guided by *technical norms* - like aggregating various kinds of fuels into a total energy input based on their individual Btu contents. In some other cases the aggregation may be *data driven* – like using Principal Component Analysis where several inputs (or outputs) are replaced by a small number of their principal components, to reduce the problem of dimensionality.[1] An *economically meaningful* aggregation procedure is one where a sub group of inputs (or outputs) is replaced by its total cost (or revenue). Using prices would be the valid way to aggregate when the relevant inputs (or outputs) have been chosen in an allocatively efficient way.

[1] For a discussion of the use of Principal Component Analysis (PCA) within DEA see Adler and Golany (2001, 2002). An inherent problem with PCA is that typically there is no intuitive or economic interpretation of what these principal components represent.

This paper extends the literature in several ways. First, we address the problem of input aggregation in DEA models within a more general framework. We show a direct link between input aggregation in DEA on one hand and the test of parameter restrictions implied by input aggregation in an explicitly specified production function on the other. Second, departing from the existing practice of assuming identical prices across firms, we allow the input prices to vary and show how the DEA LP problems for the aggregated model have to be modified to accommodate this variation. Finally, we apply the proposed method to the Indian manufacturing sector using state level data. We then carry out a statistical test of the validity of input aggregation using Banker's (1993) F test. In this case it can also serve as a test of allocative efficiency. We find that the validity of aggregating production workers and non-production workers into a composite labor input using state specific prices for aggregation cannot be rejected.

The rest of the paper is organized as follows. Section 2 provides the nonparametric DEA methodology emphasizing how the aggregate model should be set up when all firms do not face the same prices of the inputs being aggregated. Section 3 spells out the test procedure proposed by Banker (1993). Section 4 reports the empirical results. Section 5 concludes.

2. The DEA Methodology

In parametric models, one specifies an explicit functional form for the frontier and econometrically estimates the parameters using sample data for inputs and output. Hence the validity of the derived technical efficiency measures depends critically on the appropriateness of the functional form specified.

In contrast, the method of DEA introduced by Charnes, Cooper and Rhodes (CCR) (1978) and further generalized by Banker, Charnes, and Cooper (BCC) (1984) provides a nonparametric alternative to parametric frontier production function analysis. In DEA, one makes only a few fairly weak assumptions about the underlying production technology. In particular, no functional specification is necessary. Based on these assumptions a production frontier is empirically constructed using mathematical programming methods from observed input-output data of sample firms. Efficiency of firms is then measured in terms of how far they are from the frontier.

Consider an industry producing a scalar output, y, from a bundle of m inputs, $x=(x_1,x_2,\ldots,x_m)$. Let (x^j, y_j) be the observed input-output bundle of firm j $(j= 1,2,\ldots, N)$. The technology is defined by the production possibility set

$T =\{(x, y)$: y can be produced from $x \}$. (1)

The corresponding production function is

$f(x) = \max y : (x, y) \in T.$

An input-output combination (x^0, y_0) is feasible if and only if $y_0 \leq f(x^0)$. The production function evaluated at the input bundle x^0 is

$$f(x^0) = \max \varphi y_0 : (x^0, \varphi y_0) \in T.$$

Obviously, every observed input-output bundle (x^j, y_j) is feasible. Under the standard assumptions of convexity of the production possibility set and free disposability of inputs and output, an estimate of the set T is;

$$S = \left\{ (x, y) : x \geq \sum_{j=1}^{N} \lambda_j x^j ; y \leq \sum_{j=1}^{N} \lambda_j y_j ; \sum_{j=1}^{N} \lambda_j = 1; \lambda_j \geq 0; j = 1,2,...,,N \right\}. \qquad (2)$$

A nonparametric estimate of the production frontier at the input bundle x^0 is obtained by solving the following linear programming problem:

$$\begin{aligned} & \hat{f}(x^0) = \max \varphi y_0 \\ \text{s.t} \quad & \sum_{j=1}^{N} \lambda_j y_j \geq \varphi y_0 ; \\ & \sum_{j=1}^{N} \lambda_j x^j \leq x^0 ; \\ & \sum_{j=1}^{N} \lambda_j = 1; \\ & \lambda_j \geq 0; (j = 1,2,..., N). \end{aligned} \qquad (3)$$

The support of $\hat{f}(.)$ is the free disposal convex hull of the observed input bundles

$$X^* = \left\{ x : x \geq = \sum_{j=1}^{N} \lambda_j x^j ; \sum_{j=1}^{N} \lambda_j = 1; \lambda_j \geq 0; j = 1,2,..., N \right\}. \qquad (4)$$

The dual of the LP problem (3) is

$$\begin{aligned} & \min \alpha_0 + \beta^{0\prime} x^0 \\ \text{s.t.} \quad & \alpha_0 + \beta^{0\prime} x^j \geq p y_j ; \quad (j=1,2,\ldots,N); \\ & p y_0 = y_0 ; \end{aligned} \qquad (5)$$

$p \geq 0; \beta^0 \geq 0;\ \alpha_0$ unrestricted.

For a simple example, consider the 3-input 1-output case and the observed input-output bundle of firm *k* in the sample. For this example the explicit form of problem (3) above is

$$\hat{f}(x_{1k}, x_{2k}, x_{3k}) = \max \varphi y_k$$

$$\text{s.t. } \sum_{j=1}^{N} \lambda_j y_j \geq \varphi y_k;$$

$$\sum_{j=1}^{N} \lambda_j x_{1j} \leq x_{1k};$$

$$\sum_{j=1}^{N} \lambda_j x_{2j} \leq x_{2k}; \qquad (6)$$

$$\sum_{j=1}^{N} \lambda_j x_{3j} \leq x_{3k};$$

$$\sum_{j=1}^{N} \lambda_j = 1;$$

$$\lambda_j \geq 0; (j = 1,2,..., N);\ \varphi \text{ unrestricted.}$$

The corresponding dual problem is

$$\hat{f}(x_{1k}, x_{2k}, x_{3k}) = \min\ \alpha_k + \beta_1^k x_{1k} + \beta_2^k x_{2k} + \beta_3^k x_{3k}$$

$$\text{s.t.} \quad \alpha_k + \beta_1^k x_{1j} + \beta_2^k x_{2j} + \beta_3^k x_{3j} \geq p y_j; \qquad \textit{(j=1,2,…,k,…,N);} \qquad (7)$$

$$p y_k = y_k;$$

$$\beta_1^k, \beta_2^k, \beta_3^k, p \geq 0; \alpha_k \text{ unrestricted.}$$

Clearly, in (7) above, $p = 1$.

Define the function[2]

$$R^k(x_1, x_2, x_3) = \alpha_k + \beta_1^k x_1 + \beta_2^k x_2 + \beta_3^k x_3 \qquad (8)$$

Then $R^k(x^j) \geq y_j$ for each *j* and $R^k(x^k) = \hat{f}(x^k)$.

We may express the deviation between y_j and $R^k(x^j)$ as

$$\varepsilon_j^k = R^k(x^j) - y_j\ (j = 1,2,..., k,.., N).$$

[2] The function $R^k(.)$ is a supporting hyperplane constituting a local outer approximation to the frontier of the production possibility set and is one of the facets of the *over production function* (Varian, 1984). For a more detailed discussion see Ray (2004) chapter 10.

Given that $R^k(x^k) = \hat{f}(x^k)$, as noted above, then problem (7) can be reformulated as

$$\min \ \varepsilon_k^k = R^k(x^k) - y^k \tag{9}$$

s.t. $\varepsilon_j^k = R^k(x^j) - y_j \geq 0 (j = 1,2,...,k,..,N).$

Several points need to be noted here:

(i) While the non-negative residual ε_k^k measures the deviation of y_k from the frontier $\hat{f}(.)$ evaluated at the input bundle x^k, the other residuals $\varepsilon_j^k (j \neq k)$ have no such interpretation;

(ii) For any other input bundle $x^j (j \neq k)$, the corresponding tangent hyper-plane $R^j(x_1, x_2, x_3) = \alpha_j + \beta_1^j x_1 + \beta_2^j x_2 + \beta_3^j x_3$ will have a different gradient vector.

(iii) The efficiency residuals $\varepsilon_j^j (j = 1,2,...,k,..,N)$ are obtained independently rather than jointly.

In (7) or (9) above, we impose no additional constraints beyond the non-negativity of the βs. Now impose a further constraint

$$a\beta_1^k - b\beta_2^k = 0. \tag{10}$$

That is, $\beta_2^k = \frac{a}{b}\beta_1^k$. The restricted version of (7) would then be

$$\tilde{R}(x^k) = \min \alpha_k + \beta_1^k (x_{1k} + \tfrac{a}{b} x_{2k}) + \beta_3^k x_{3k}$$

s.t. $\alpha_k + \beta_1^k (x_{1j} + \tfrac{a}{b} x_{2j}) + \beta_3^k x_{3j} \geq p y_j$; *(j=1,2,…,k,…,N);* (11)

$$p y_k = y_k;$$

$\beta_1^k, \beta_2^k, \beta_3^k, p \geq 0; \alpha_k$ unrestricted.

Define, now, the aggregated input

$$X_{1j} = x_{1j} + \tfrac{a}{b} x_{2j}. \tag{12}$$

Problem (11) would then become

$$\tilde{R}(x^k) = \min \ \alpha_k + \beta_1^k X_{1k} + \beta_3^k x_{3k}$$

s.t. $\alpha_k + \beta_1^k X_{1j} + \beta_3^k x_{3j} \geq p y^j$; *(j=1,2,…,k,…,N);* (13)

$$p y_k = y_k;$$

$\beta_1^k, \beta_3^k, p \geq 0; \alpha_k$ unrestricted.

The dual of this problem is

$$\tilde{f}^k(x^k) = \max \varphi y_k$$

$$\text{s.t.} \sum_{j=1}^{N} \lambda_j y^j \geq \varphi y_k;$$

$$\sum_{j=1}^{N} \lambda_j X_{1j} \leq X_{1k};$$

$$\sum_{j=1}^{N} \lambda_j x_{3j} \leq x_{3k}; \qquad (14)$$

$$\sum_{j=1}^{N} \lambda_j = 1;$$

$$\lambda_j \geq 0; (j = 1,2,..., N).$$

Obviously, when $a = b$, X_1 is simply the sum of the quantities of the inputs x_1 and x_2. In that case, the two inputs are treated as perfect substitutes. Further, because (13) is a restricted version of (7), the minimum value of the objective function at the optimal solution of (13) will be no lower than what is obtained at the optimal solution of (7). Therefore,

$$\tilde{\varepsilon}_k^k = \tilde{f}^k(x^k) - y_k \geq \varepsilon_k^k = \hat{f}(x^k) - y_k. \qquad (15)$$

The former is the restricted and the latter the unrestricted residual.

In econometric models, test of linear restrictions on parameters are usually conducted by comparing the residual sums of squares from the restricted and the unrestricted model in an appropriate F test. A comparable F test developed by Banker (1993) can be employed to test whether the efficiency residuals from the restricted DEA model are significantly bigger as a whole than the unrestricted DEA residuals.[3] Note that the restriction (10) results in the input aggregation defined in (12). Hence, this F test is in effect a test of the validity of the way the inputs have been aggregated.

The multipliers β_1^k and β_2^k at the optimal solution of (7) denote the shadow prices of the corresponding inputs of firm k: x_1^k and x_2^k. A problem often encountered in DEA (as in other linear programming models) is that at the optimal solution, the shadow price becomes zero because a constraint is non-binding. To avoid this problem one may impose upper and lower bounds of the form

[3] Pastor, Ruiz, and Sirvent (1995) performed a nonparametric statistical test of nested radial DEA models to determine the optimal choice of inputs and outputs.

$$L \leq \frac{\beta_2^k}{\beta_1^k} \leq U$$

on the ratio of a pair of shadow prices. This is known as assurance region (AR) analysis. Note that one gets the restriction (10) above by setting

$$L = \tfrac{a}{b} = U.$$

In that sense, aggregation of the two inputs in the manner described above is a degenerate case of AR analysis.

The lower and upper bounds in AR analysis are often arbitrarily chosen. However, the cost minimizing behavior of a competitive firm can provide guidance in selecting the ratio $\frac{a}{b}$ from available data. When firm *k* is allocatively efficient in its choice of the input quantities x_1^k and x_2^k, the marginal rate of substitution will be equal to the ratio of the two input prices faced by the firm. In that case,

$$\frac{\beta_2^k}{\beta_1^k} = \frac{w_2^k}{w_1^k}.$$

A valid measure of the aggregate input would then be[4]

$$X_{1j} = x_{1j} + \tfrac{w_2^k}{w_1^k} x_{2j}.$$

As argued by Banker et al. (2005), Banker's F test would then become a test of allocative efficiency.

In the existing literature it is standard practice to assume that all firms face the same prices of the inputs. In that case

$$\frac{w_2^k}{w_1^k} = \frac{\overline{w}_2}{\overline{w}_1} \text{ for each firm } k.$$

When this is not the case, one must construct a different measure of the aggregate input by using a different price ratio in order to evaluate the efficiency of each firm *k*. That is when evaluating firm *r* the aggregate input is constructed as

$$X_{1j}^r = x_{1j} + \tfrac{w_2^r}{w_1^r} x_{2j}.$$

However, when evaluating a different firm, *s*, the appropriate way to measure the aggregate input would be to define

[4] In aggregating over more than two inputs, any one input price may be regarded as a *numeraire* and the other inputs may be weighted by their relative prices.

$$X_{1j}^{s} = x_{1j} + \frac{w_2^s}{w_1^s} x_{2j} \; (j = 1,2,..,N).$$

The relevant LP problem (12) for any firm k ($k=1,2,\ldots,N$) would then become

$$\tilde{f}^{k}(x^{k}) = \max \varphi y_k$$

$$\text{s.t.} \sum_{j=1}^{N} \lambda_j y_j \geq \varphi y_k;$$

$$\sum_{j=1}^{N} \lambda_j X_{1j}^{k} \leq X_{1k}^{k};$$

$$\sum_{j=1}^{N} \lambda_j x_{3j} \leq x_{3k}; \qquad (16)$$

$$\sum_{j=1}^{N} \lambda_j = 1;$$

$$\lambda_j \geq 0; (j = 1,2,..., N).$$

Note that the aggregate input X_1 needs to be measured differently while solving the problem (16) for each individual firm unless all firms face the same prices of the two inputs.

It is possible, of course, that in a particular application input (or output) prices are not available and one must use the actual cost (or revenue) data in the model, as in the case of Banker (2005). This, it may be noted, amounts to assuming that input prices are identical across firms.

3. DEA as Maximum Likelihood Estimation and Banker's F Test

We start with N observed input-output bundles. The pair (x^j, y_j) represents the input bundle x^j used by firm j to produce the scalar output y_j. Next, following Banker (1993), consider the production function mapping from the n-element input bundle $x^0 \in X \subseteq R_+^n$ onto the non-negative scalar output y_0:

$$y_0 = g(x^0). \qquad (17)$$

We assume that the production function satisfies the following postulates:.

(P1) $g(x)$ is monotonic in x. That is if $x^{1'} \geq x^2$, then $g(x^1) \geq g(x^2)$.

(P2) $g(x)$ is concave. Hence, if $x^1, x^2 \in X$ and $x^* = \lambda x^1 + (1-\lambda)x^2$, $0 < \lambda < 1$, then

$g(x^*) \geq \lambda\, g(x^1) + (1-\lambda)\, g(x^2)$.

(P3) For each observation (x^j, y_j), $g(x^j) \geq y_j$; $(j = 1,2,\ldots, N)$.

(P4) For *any* other function $\tilde{g}(x)$ also satisfying (P1-P3), $\tilde{g}(x) \geq g(x)$ for all $x \in X$.

Now consider the set $X^* = \{x : x \geq \sum_{j=1}^{N} \lambda_j x^j; \sum_{j=1}^{N} \lambda_j = 1; \lambda_j \geq 0\} \subseteq X$. Clearly, X^* is the free disposal convex hull of the observed input bundles. Banker has shown that the unique function $y = g(x)$ determined for $x \in X^*$ by the postulates (P1-P4) corresponds to that estimated by DEA.

An implication of this is that the deviation $\varepsilon_j = \tilde{g}(x^j) - y_j$ is minimized for each observation *j* by the function $g^*(x)$.

Now consider the frontier production function

$$y = g(x) - \varepsilon; \varepsilon \geq 0. \tag{18}$$

Here, the non-negative deviation of the observed output *y* from the frontier *g(x)* has some one-sided probability distribution *f(ε)*. Then the likelihood maximization problem can be specified as:

$$\underset{f(.),g(.)}{\text{maximize}}\, L = \prod_{j=1}^{N} f\left(\varepsilon_j = g(x^j) - y_j\right)$$

$$\text{subject to} \quad g(x^j) - y_j \geq 0; \tag{19}$$

$g(.)$ is a monotone increasing and concave function.

It may be noted that the DEA efficiency residuals ε_j are obtained independently of each other. This is in contrast with the frontier production function model proposed by Aigner and Chu (1968). In their case, a single parametric function is fitted to the entire data set and the efficiency residuals are jointly derived and, therefore, are not independent of one another. Now suppose that we choose a probability density function *f(.)* such that $f(\varepsilon_j)$ is monotone decreasing in the efficiency residuals. In that case, because the DEA estimate of the production function minimizes each ε_j, it thereby maximizes each $f(\varepsilon_j)$. Hence, the DEA frontier $g^*(x)$ maximizes the likelihood function subject to the constraints specified above.

Banker specifies the deterministic frontier where the random inefficiency component of *y* appears in an additive manner. One may directly link the one-sided econometric frontier with the DEA frontier by specifying (18) differently as

$$y = g(x)e^{-\varepsilon}; \varepsilon \geq 0 \tag{20}$$

leading to

$$g(x) = \varphi y; \varphi \geq 1. \tag{21}$$

Thus,

$$\varepsilon = \ln(\varphi) \geq 0. \tag{22}$$

Note that all the preceding arguments about the DEA frontier $g^*(x)$ as a maximum likelihood estimator of the unknown frontier $g(x)$ remains valid.

An alternative conceptualization of the frontier would be to directly specify (21) with the non-negative efficiency residual defined as

$$\varepsilon = (\varphi - 1) \geq 0. \tag{23}$$

Banker has proposed a number of statistical tests for comparing two groups of firms to assess whether one group is more efficient than the other. Assume that there are N firms in the sample of which m_1 are in group 1 and m_2 are in group 2. Firms in group1 have the exponential distribution of (in)efficiency ε_j with parameter σ_1 and those in group 2 also have the exponential distribution but with parameter σ_2 . Designate the first group of firms as M_1 and the second group as M_2. Consider the residuals $\varepsilon_j^*(j = 1, \ldots, N)$ obtained from DEA. Under the maintained hypothesis, the sample statistic

$\sum_{j \in M_i} \frac{\varepsilon_j^*}{\sigma_i}$ has the χ^2 distribution with $2m_i$ $(i = 1, 2)$ degrees of freedom.

Under the null hypothesis $\sigma_1 = \sigma_2$, the test statistic

$$F = \frac{\sum_{j \in M_1} \varepsilon_j^* \Big/ m_1}{\sum_{j \in M_2} \varepsilon_j^* \Big/ m_2} \tag{24}$$

has the F distribution with $(2m_1, 2m_2)$ degrees of freedom.

On the other hand, if the ε_j s have the half Normal distribution, (i.e., the Normal distribution with mean 0 and variance σ^2 truncated from below at 0), then $\sum_{j \in M_1} \left(\frac{\varepsilon_j^*}{\sigma_1} \right)^2$ has the χ^2 distribution with m_1 degrees of freedom. Similarly, $\sum_{j \in M_2} \left(\frac{\varepsilon_j^*}{\sigma_2} \right)^2$ has the χ^2 distribution with m_2 degrees of freedom. Hence, in this case, under the null hypothesis $\sigma_1 = \sigma_2$, the statistic

$$F = \frac{\sum_{j \in M_1} (\varepsilon_j^*)^2 \Big/ m_1}{\sum_{j \in M_2} (\varepsilon_j^*)^2 \Big/ m_2} \tag{25}$$

has the F distribution with (m_1, m_2) degrees of freedom.

One would use $\varepsilon_k^* \equiv \ln(\hat{\varphi}_k)$ for the aggregated (i.e., the restricted) model in the numerator and $\varepsilon_k^* \equiv \ln(\varphi_k)$ for the disaggregated (i.e., the unrestricted) model in the denominator in testing for the validity of input aggregation. The aggregated model would be rejected only when the test statistic exceeds the critical value for the relevant degrees of freedom, for any specific distributional assumption.

4. Application to Indian Manufacturing

In the empirical application we use state level aggregate data on output and inputs in total manufacturing for the different states (and union territories) of India from the Annual Survey of Industries (ASI) for the year 2003-04. A single output *(y)* measured by the value of production at current prices is considered. Because it is a single cross section data set and state level indexes of output price are not available, the quantity of output is treated as proportional to its value. Inputs considered were - production workers *(L_1)*, non-production workers *(L_2)*, fixed capital *(K)*, fuels *(F)*, and materials *(M)*. The two labor inputs are measured in numbers of persons employed. All other inputs are expressed in value terms. The data used are reported in Table 1.

Two different DEA models were considered. In one, the two labor inputs are treated separately and the LP model similar to (6) was used. In the other, they are combined into a single labor input *(TL)* and LP model similar to (16) was utilized. Table 1 reports the own-price weighted aggregate labor input for each state. As mentioned earlier, when evaluating the efficiency of a particular state using the LP model (16), the price weights of the two types of labor of that state were used for constructing the aggregate labor input for all states for that specific LP problem. There was considerable variation in the prices and also in the price ratios of the two separate labor inputs across states. Using all-India data, we found that the annual earnings of a non-production worker were 3.120 times the earnings of a production worker. At the state level, this same ratio ranged from a low of 1.715 in case of Jharkhand (JH) to a high of 5.373 in case of Chattisgarh (CT). For comparison, Table 2 reports the aggregate labor constructed using these three different price ratios. We not only find different figures for aggregate labor based on these three different price ratios but that the ordering of the states in terms of the aggregate labor input also changes. To take one example, the size of the aggregate labor input is smaller in Himachal Pradesh (HP) than in Jammu and Kashmir (JK) based on the all-India prices and also based on the prices prevailing in Jharkhand (JK) but is larger than that of Jammu and Kashmir based on the prices of Chattisgarh. When there are differences in input prices across observations,

using a common set of prices for all observations may not represent an accurate method for aggregating inputs.

Since we are using state level data, we assume that constant returns to scale holds.[5] As such, the constraint pertaining to the sum of λ_js was removed from the DEA models. The optimal DEA objective function values and the corresponding (inefficiency) residuals obtained from (22) are reported in Table 3. The columns labeled φ_k and ε_k relate to the disaggregated model where the two kinds of labor are treated as two distinct inputs. Similarly, $\hat{\varphi}_k$ and $\hat{\varepsilon}_k$ relate to the model where the aggregated labor is treated as a single input. As expected, for many states, DEA inefficiency residuals are larger in the restricted model with aggregated labor. The summary statistics relevant for the F tests are reported in Table 4. Under the assumption that ε has an exponential distribution, the test statistic is

$$F = \frac{2.076605}{1.5287587} = 1.35836 .$$

The relevant p-value from the F distribution with 46 degrees of freedom for both the numerator and the denominator is 0.15126 in a 1-tailed test. Thus the model using aggregate labor as one input cannot be rejected.

For the alternative assumption that ε has a half-Normal distribution, the test statistic is

$$F = \frac{0.2932929}{0.1737017} = 1.688486.$$

The p-value from $F_{23,23}$ for a 1-tailed test is 0.1083. Again we are unable to reject the aggregated model under the half-Normal distributional assumption.

In the above tests, we performed log transformation of the φs to obtain the inefficiencies. Since $\varphi \geq 1$, such a transformation always yields a non-negative value. Using the transformation as in (23) above we performed another set of F-tests. This time the respective F statistics for the exponential and the half-Normal assumptions were 1.378187 and 1.764355. Again, the null hypothesis that the aggregation is valid cannot be rejected.[6]

[5] While the observed individual firm level output can be produced from the firm level input bundles, the aggregate state level input- output combination can be considered feasible only under the assumption of additivity, i.e., under constant returns to scale.

[6] For any F-distribution with equal degrees of freedom for the numerator and the denominator the median equals 1. Therefore, in the present context, the relevant p-value for the F-distribution truncated from below at 1 will be twice the tabulated value. Hence the null hypothesis cannot be rejected *a fortiori*.

5. Conclusion

A problem frequently encountered in DEA is that the total number of inputs and outputs included tend to be too many relative to the sample size. Because each additional input or output included in the analysis imposes another constraint, the set of feasible solutions tends to become smaller and more and more firms tend to lie on or close to the frontier. One way to counter this problem is to combine several inputs (or outputs) into (meaningful) aggregate variables reducing thereby the dimension of the input (or output) vector. In this paper we propose that own-price-weights should be used to aggregate inputs for individual observations and show how an F test developed by Banker may be applied under appropriate distributional assumptions of the efficiency component to test the validity of input aggregation. The empirical application using data from Indian manufacturing suggests that using own-price aggregated measure of total labor is a valid aggregation of production and non-production workers in this specific context. This, in its turn implies that the choice of production and non-production workers by firms in the individual states is consistent with sub-vector allocative efficiency.

Finally, the findings reported in this paper are based on the two explicit specifications of the one-sided efficiency term – the half-Normal and the exponential distributions. For a more general distribution free test of the aggregation restriction one may examine the quantiles of the empirical density function of the test statistic obtained through smoothed bootstrap.

References

Adler, N. and B. Golany (2001) "Evaluation of Deregulated Airline Networks using Data Envelopment Analysis with Principal Component Analysis with an Application to Western Europe" *European Journal of Operational Research*, 132, 260-273.

Adler, N. and B. Golany (2002) "Including Principal Component Weights to Improve Discrimination in Data Envelopment Analysis", *Journal of Operational Research Society*, 53, 985-991.

Aigner, D.J. and S. F.Chu (1968), "On Estimating the Industry Production Function," *American Economic Review* 58:4 (September) 826-39.

Banker, R. D. (1993), "Maximum Likelihood, Consistency, and Data Envelopment Analysis: A Statistical Foundation", *Management Science,* 39, 1265-1273.

Banker, R. D., H. Chang, and R. Natarajan (2005), "Estimating DEA Technical and Allocative Inefficiency Using Aggregate Cost or Revenue Data" working paper.

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

Charnes, A., W.W. Cooper, and E. Rhodes (1978) "Measuring the Efficiency of Decision Making Units," *European Journal of Operational Research* 2:6 (November), 429-44.

Färe, R., S. Grosskopf, and V. Zalenyuk (2004) "Aggregation Bias and its Bounds in Measuring Technical Efficiency" *Applied Economics Letters*, 11, 657-660.

Färe, R. and V. Zelenyuk (2002) "Input Aggregation and Technical Efficiency" *Applied Economics Letters*, 9, 635-636.

Government of India (2005) *Annual Survey of Industries 2003-04.*

Leibenstein, H. and S. Maital (1992) "Empirical Estimation and Partitioning of X-Inefficiency: A Data-Envelopment Approach" *The American Economic Review*, 82, 2, 428-433.

Pastor, J.T., J.J. Ruiz, and I. Sirvent (1995) "A Statistical Test for Nested Radial DEA Models"; *Working Paper*, Departmento de Estadistical e Investigacion Operativa; Universidad de Alicante, Spain.

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

Simar, L. and P. W. Wilson (2001) "Testing Restrictions in Nonparametric Efficiency Models" *Communications in Statistics: Simulation and Computation*, 30, 1, 159-184.

Sirvent, I., J. L. Ruiz, F. Borrás, and J. T. Pastor (2005) "A Monte Carlo Evaluation of Several Tests for the Selection of Variables in DEA Models" *International Journal of Information Technology and Decision Making*, 4, No. 3, 325-344.

Tauer, L.W. (2001) “Input Aggregation and Computed Technical Efficiency” *Applied Economics Letters*, 8, 295-297.

Thomas, A.C. and L.W. Tauer (1994) “Linear Input Aggregation Bias in Nonparametric Technical Efficiency Measurement” *Canadian Journal of Agricultural Economics*, 42, 77-86.

Varian, H. R.(1984), “The Nonparametric Approach to Production Analysis,” *Econometrica* 52:3 (May) 579-97.

Table 1: Input and output data used in the DEA models.

N	State	K	L1	L2	TL	F	M	Y
1	**JK**	111.6637	64.30702	14.5	96.86611	26.61988	371.7456	583.6784
2	**HP**	1078.081	52.1434	17.20189	119.801	109.0358	972.7208	1642.979
3	**PU**	135.0711	38.61039	10.47716	64.0258	44.63023	415.2789	641.0619
4	**CH**	118.5856	21.14449	12.8403	51.58124	24.19011	300.7719	541.346
5	**UT**	321.3196	40.63623	20.5729	79.81679	73.67599	615.7614	1067.571
6	**HA**	354.8448	55.05838	19.56436	121.4415	53.79601	937.1566	1452.788
7	**DE**	65.84485	24.84079	11.27995	56.17386	14.4645	287.6334	499.4051
8	**RA**	257.0024	35.14105	9.846845	64.47018	68.80301	392.2278	690.3072
9	**UP**	323.9802	47.55516	14.11021	90.79393	59.50633	643.6419	1011.158
10	**BI**	207.726	32.04452	7.273288	50.6761	48.79521	468.7178	608.0212
11	**AS**	426.4707	61.10573	11.50127	103.5866	44.42548	629.7261	1038.613
12	**WB**	405.4254	68.82161	17.89448	108.1636	52.63649	542.0813	879.0629
13	**JH**	1124.924	76.76088	23.12094	116.402	178.548	868.3324	1922.115
14	**OR**	960.3772	58.65912	15.8242	106.6652	150.0763	561.1186	1102.566
15	**CT**	655.2008	55.29266	21.74517	172.1244	213.5977	703.9776	1576.258
16	**MP**	464.5728	53.98256	17.8283	107.4929	112.0617	886.508	1433.411
17	**GU**	670.4852	42.05393	14.94568	87.71353	78.64564	1102.93	1620.432
18	**MA**	477.6942	44.29301	19.46286	93.76306	64.06032	788.8395	1363.702
19	**AP**	231.1586	48.82516	9.552898	82.51308	33.46811	347.6823	558.0892
20	**KA**	501.3298	54.93378	16.86614	116.2566	56.82595	689.1845	1143.353
21	**GO**	681.0273	45.5592	17.20401	89.80894	80.85246	1255.872	2116.574
22	**KE**	126.198	49.61446	8.045529	81.59947	32.73065	395.8028	579.7103
23	**TN**	229.2859	46.80381	10.61958	83.43632	38.83775	376.7324	641.7911

Notes:

(i) **Variable Definition:** K = fixed capital; L_1 = production workers; L_2 = non-production workers; TL = total (aggregated) labor; F = fuel; M = materials; Y = output.

Source: ASI 2003-04; All variables are expressed per-factory; Labor inputs are measured by the number of persons employed, all other inputs are in *lakhs* of Rupees at current prices (1 *lakh* = 0.1 million).

(ii) **State Names:**

JK: Jammu & Kashmir; HP: Himachal Pradesh; PU: Punjab; CH: Chandigarh; UT: Uttaranchal; HA: Haryana; DE: Delhi; RA: Rajasthan; UP: Uttar Pradesh; BI: Bihar; AS: Assam; WB: West Bengal; JH: Jharkhand; OR: Orissa; CT: Chhattisgarh; MP: Madhya Pradesh; GU: Gujarat; MA: Maharashtra; AP: Andhra Pradesh; KA: Karnataka; GO: Goa; KE: Kerala; TN: Tamil Nadu.

Table 2: Labor input aggregation based on all-India prices and prices of selected states

N	State	L1	L2	TL based on CT prices	TL based on JH prices	TL based on All India prices
1	**JK**	64.3070	14.5000	142.2122	89.1674	109.5502
2	**HP**	52.1434	17.2019	144.5652	81.6362	105.8171
3	**PU**	38.6104	10.4772	94.9018	56.5736	71.3014
4	**CH**	21.1445	12.8403	90.1325	43.1593	61.2091
5	**UT**	40.6362	20.5729	151.1697	75.9087	104.8282
6	**HA**	55.0584	19.5644	160.1732	88.6017	116.1035
7	**DE**	24.8408	11.2799	85.4453	44.1804	60.0367
8	**RA**	35.1410	9.8468	88.0459	52.0236	65.8654
9	**UP**	47.5552	14.1102	123.3660	71.7473	91.5821
10	**BI**	32.0445	7.2733	71.1222	44.5147	54.7388
11	**AS**	61.1057	11.5013	122.8994	80.8248	96.9922
12	**WB**	68.8216	17.8945	164.9645	99.5019	124.6563
13	**JH**	76.7609	23.1209	200.9843	116.4020	148.9033
14	**OR**	58.6591	15.8242	143.6789	85.7899	108.0341
15	**CT**	55.2927	21.7452	172.1244	92.5750	123.1424
16	**MP**	53.9826	17.8283	149.7699	84.5494	109.6108
17	**GU**	42.0539	14.9457	122.3536	67.6785	88.6878
18	**MA**	44.2930	19.4629	148.8624	77.6623	105.0214
19	**AP**	48.8252	9.5529	100.1507	65.2037	78.6323
20	**KA**	54.9338	16.8661	145.5516	83.8510	107.5598
21	**GO**	45.5592	17.2040	137.9923	75.0557	99.2395
22	**KE**	49.6145	8.0455	92.8412	63.4086	74.7183
23	**TN**	46.8038	10.6196	103.8604	65.0112	79.9392

Table 3: DEA Results from Aggregated and Disaggregated Models

N	State	$\hat{\varphi}_k$	φ_k	$\hat{\varepsilon}_k$	ε_k
1	**JK**	1.13488	1.13488	0.126527	0.126527
2	**HP**	1.09804	1.09421	0.093527	0.090033
3	**PU**	1.12736	1.02298	0.119879	0.02272
4	**CH**	1.00745	1	0.007422	0
5	**UT**	1.07424	1.07184	0.071613	0.069377
6	**HA**	1.01638	1	0.016247	0
7	**DE**	1	1	0	0
8	**RA**	1.12453	1.09509	0.117365	0.090837
9	**UP**	1.13992	1.11746	0.130958	0.111058
10	**BI**	1.31212	1.1904	0.271644	0.174289
11	**AS**	1.06566	1.04521	0.063594	0.044218
12	**WB**	1.15938	1.15363	0.147885	0.142913
13	**JH**	1	1	0	0
14	**OR**	1.13474	1.13133	0.126404	0.123394
15	**CT**	1	1	0	0
16	**MP**	1.09964	1.08265	0.094983	0.079412
17	**GU**	1.15268	1.13472	0.14209	0.126386
18	**MA**	1.02958	1.02633	0.029151	0.025989
19	**AP**	1.16967	1.15037	0.156722	0.140084
20	**KA**	1.08521	1.07787	0.081774	0.074987
21	**GO**	1	1	0	0
22	**KE**	1.18588	1	0.170485	0
23	**TN**	1.11443	1.09039	0.108343	0.086535

Notes: $\hat{\varphi}_k$ and $\hat{\varepsilon}_k$ are from the model using aggregated labor input; φ_k and ε_k are from the model using disaggregated labor inputs.

Table 4: Results from Banker's F-tests

Based on Log Transformation		
Distribution	**F-test statistic**	**p-value based on F(23, 23)**
Half-Normal	1.688486	0.1083
Distribution	**F-test statistic**	**p-value based on F(46, 46)**
Exponential	1.35836	0.15126
Based on φ -1 Transformation		
Distribution	**F-test statistic**	**p-value based on F(23,23)**
Half-Normal	1.764355	0.090448
Distribution	**F-test statistic**	**p-value based on F(46, 46)**
Exponential	1.378187	0.140162