

Centre for Efficiency and Productivity Analysis

Working Paper Series
No. WP04/2014

Cost Constrained Industry Inefficiency
Antonio Peyrache

Date: 30th April 2014

School of Economics
University of Queensland
St. Lucia, Qld. 4072
Australia

ISSN No. 1932 - 4398

Cost Constrained Industry Inefficiency

By Antonio Peyrache

Draft: April 30, 2014

In this paper a definition of industry inefficiency in cost constrained production environments is introduced. This definition uses the indirect directional distance ducntion and quantifies the inefficiency of the industry in terms of the overall output loss, given the industry cost budget. The industry inefficiency indicator is then decomposed into sources components: reallocation inefficiency arising from sub-optimal configuration of the industry; firm inefficiency arising from a failure to select optimal input quantities (given the prevalent inputs prices); firm inefficiency due to lack of best practices. These notions are then extended to the dynamic case in order to define the industry productivity change and its associated decomposition. The method is illustrated using data on Ontario electricity distributors. These data show that lack of best practices is only a minor component of the overall inefficiency of the industry (less than 10%), with reallocation inefficiency accounting for more than 75% of the overall inefficiency of the system. A counterfactual analysis based on counterfactual input prices is conducted in order to illustrate how the model can be used to estimate the effects of a change in the regulation regime.

JEL: *D24, L23, M11*

Keywords: *Industry Inefficiency, DEA, Indirect Directional Distance Function, Productivity, optimal configuration.*

In 1957 Farrell proposed an analytical method to quantify the production inefficiency of firms (Farrell, 1957). The ideas presented in this seminal paper underpinned the very core of much of the work that has been done in the last 30 years to measure and explain efficiency and productivity. Though a lot has been written on this subject, there is still one gap to fill, which concerns the definition and quantification of efficiency and productivity measures at the industry level. This distinction between the firm and the industry level was already quite sharp and clear in Farrell (1957). The sense in which this gap must be filled is very specific and can be summarized in two points.

The first point has to do with the nature of the industries in which we want to measure inefficiency. In fact, since economists generally have a market oriented vision of the world, the economic discipline has been mainly oriented towards the measurement of efficiency in market oriented industries (such as agriculture or manufacturing). These can be defined as industries in which it is possible to identify (with reasonable levels of detail) markets for inputs and outputs; and this implies that the data structures with which one has to deal usually involve input and output *quantieties* as well as input and output *prices*. It is unfortunate that these methods cannot directly be extended to public sector industries (such as education, health, transport, energy, etc.). The reason behind such a difficulty is that in these type of industries a market for the inputs can still be identified, but a market for the outputs is very difficult to define. Therefore in such industries it is not uncommon to face datasets with input and output quantities and input prices only. For example it is possible to define the cost of running a hospital (with the associated shares for each input used), it is possible to define the output of the hospital (within reasonable limits), but it is quite hard to identify a vector of prices for these outputs. Another important feature of these industries, which distiguish them sharply from private sector industries, is that they work most of the time according to a cost constrained behaviour. In other words, it is not uncommon for these industries to have a central planner (or some policy rule which acts as a central planner) that establishes how much cost has to be allocated to each firm or production unit. Then, once this cost is allocated, the firm is supposed to squeeze as much output as possible from the available cost resources.

The second pointhas already been mentioned at the beginning of this introduction and it regards the distinction between the firm and the industry levels. This difference was identified in Farrell (1957) in the fact that "the technical efficiency of an industry will tend to be somewhat less than the weighted average of the technical efficiencies of its constituent firms". In other words firm full efficiency is a necessary but not sufficient condition for industry efficiency. In actual fact, an industry can be inefficiently organized even if all the firms which operates in the industry are fully efficient with respect to some benchmark potential. A good example of this can be provided with a scale economy argument (though this is not the only example of inefficient organization): "It may be worth noting that whereas in the case of firms, if economies or diseconomies of scale are present, it will usually be desirable to allow for them (...), where industries are concerned this will not be so. The extent to which its firms are of optimum size may legitimately be considered part of an industry's efficiency" (Farrell, 1957). The distinction between the firm and the industry level is important because it demarkates between two different types of inefficiencies: best practice inefficiency (which is the inefficiency at firm level) and structural inefficiency (which arises from the way the industry is organized). It is the aim of this paper to shed some light on this distinction in a clear way and incorporate the idea of cost constrained operation while maintaining this distinction.

The general idea of this study (and its range of empirical applicability) can be summarized in the following terms. Assume that we observe data on the inputs, the prices of the inputs and the outputs produced by each firm. Then the methodology presented in this paper is able to provide information on the proportion of total inefficiency attributable to lack of best practice (firm inefficiency) and the proportion that can be accounted for by organizational failure at the industry level. If data of this sort are available, it is possible to determine the overall cost budget of the industry by looking at the inputs used and the input prices faced by each firm. The problem that the central planner (or a market) now faces is how to allocate this overall budget accross the different production units (given a certain number of constraints) in order to make the most product out of it. This problem corresponds to the implicit determination of the optimal structure of the industry via the determination of the optimal number of firms

that should populate the industry and the optimal allocation of resources accross these firms. Once this optimization problem is solved, one is able to determine inefficiency in terms of the directional distance function (DDF), where the inefficiency is measured in terms of the overall output loss due to inefficiencies arising from different sources: *i)* the inefficiency of the firms actually operating in the industry (lack of best practice) and *ii)* the inefficiency arising from a sub-optimal configuration of the industry. It should be mentioned here that this research methodology provides two different tools. The first one can be classified as a *benchmarking* tool and it consists in the determination of the optimal configuration of the industry. In this sense the associated industry inefficiency measure has the status of a *potential*, signalling how much production is lost due to deviations from the optimal. The second tool has the nature of a *descriptive* tool because it allows the monitoring of the performance of the industry in time. From this alternative perspective the purpose of the analysis is to look at how much the industry is *actually* improving.

The method introduced in this paper is illustrated using data on Ontario electricity distributors. The data show that lack of best practices at the firm level is only a minor component of the overall inefficiency of the industry (less than 10%). The bulk of the industry inefficiency is accounted for by severe deviations from the optimal configuration. In this empirical part a counterfactual analysis is also provided. The Ontario energy sector has been the subject of heavy regulation and subsidies for many years. Some of these subsidies impose a ceiling on the returns on capital assets, thus distorting the input prices faced by producer and inducing an overcapitalization effect. On the top of this, there is also some aneddotical evidence that, thanks to the structure of the subsidies, worker unions in the energy sector where able to bargain higher salaries with respect to other industries. The overall effect is that the relative price of the inputs is changed and accordingly the optimal input quantities are affected as well. In order to run a counterfactual analysis, it is assumed that loosening the structure of subsidies will induce an increase of the return on capital assets of about 70% and reduce salaries of about 10%. In this scenario the following changes are observed with respect to the current situation: i) a reduction in optimal input quantities, with capital showing the highest contraction at 6.1%; ii) optimal input cost shares change drastically with the capital input

share increasing from 14.6% to 22.1%; iii) a reduction of about 6.7% in the potential output of the industry. Though this reduction in the output of the industry may be considered undesirable, one must pay attention to the fact that the industry is far from being at its potential. In fact, this adverse reduction in the potential output of the industry can be easily offset by an increase in the efficiency with which the industry is operating. Since the proportion of potential output loss to the current inefficiency of the industry is about 12%, this means that in the counterfactual scenario what is decreasing is the potential, not the actual production.

The rest of the paper is organized as follows. Seciton 1 presents the defintion of firm technology and inefficiency. Section 2 extends these notions to the industry level. Section 3 introduces the notion of industry productivity change, providing a tool to benchmark efficiency gains in time. Section 4 is dedicated to the empirical illustration. Finally section 5 concludes.

I. Firm Technology and Inefficiency

Consider an industry where $\mathbf{x} \in R_+^N$ inputs produce $\mathbf{y} \in R_+^M$ outputs. It is assumed that data on inputs and outputs are available for a number of time periods $t = 1, ..., T$ and in each time period they are observed for a number K^t of firms ($k = 1, ..., K^t$). This specification allows for an unbalanced panel dataset (with the number of firms changing from one period to the next). In each time period the data can be collected into two matrices: an input matrix $\mathbf{X}^t = \left[\mathbf{x}_1^t, \ldots, \mathbf{x}_{K^t}^t\right]'$ and an output matrix $\mathbf{Y}^t = \left[\mathbf{y}_1^t, \ldots, \mathbf{y}_{K^t}^t\right]'$, where the observations relative to each firm are collected on the rows of these matrices. The dataset is the collection of these matrices accross all the time periods (note that these matrices have different dimensionality in different time periods):

$$\left(\mathbf{X}^t, \mathbf{Y}^t\right), \quad t = 1, \ldots, T \tag{1}$$

The data generated firm technology set (production set, or production possibilities set) at time t is given by the data envelopement (DEA) of the observations available in that time

period (Banker et al., 1984, 2005):

$$\Psi^t = \left\{ (\mathbf{x}, \mathbf{y}) : \lambda \mathbf{X}^t \leq \mathbf{x}' , \ \lambda \mathbf{Y}^t \geq \mathbf{y}' , \ \sum_k \lambda_k = 1 , \ \lambda \geq \mathbf{0} \right\} \tag{2}$$

This production set is built under the assumption that the technology satisfies convexity and free disposability of inputs and outputs. The constraint on the intensity vector $\sum \lambda_k = 1$ means that the technology allows for variable returns to scale (VRS). An equivalent representation of the technology is given by the output sets, which are the collection of all producible outputs given a certain input vector $P^t(\mathbf{x}) = \left\{ \mathbf{y} : (\mathbf{x}, \mathbf{y}) \in \Psi^t \right\}$ and can be represented in functional form via the directional output distance function (DDF) (see Chambers et al., 1996, 1998):

$$TE = D^t(\mathbf{x}, \mathbf{y}; \mathbf{g}_y) = \sup_{\beta} \{ \beta : (\mathbf{y} + \beta \mathbf{g}_y) \in P(\mathbf{x}) \} \tag{3}$$

The DDF provides a measure of technical efficiency (TE) or, more precisely, of technical inefficiency because it represents the total loss in output due to inefficient use of the inputs available to the firm (with respect to the benchmark represented by technology (2)). This measure of inefficiency is expressed in terms of the *numeraire* $\mathbf{g}_y$ which is therefore assumed to be fixed accross firms and time periods (comparisons of technical inefficiency using different *numeraires* would be equivalent to comparing apples with oranges). The function (3) can also be interpreted as a shortage function, inasmuch it is a measure of total output loss with respect to a potential output benchmark.

It is now interesting to consider an alternative representation of the technology (2) which assumes availability of information on input prices in the form of a row vector $\mathbf{w} \in \mathbb{R}^N_+$. In this case, one may think of the production possibilities as the collection of all the possible output vectors which are feasible when the cost budget is set at level C. This gives rise to the indirect output sets $IP(\mathbf{w}/C) = \left\{ \mathbf{y} : (\mathbf{x}, \mathbf{y}) \in \Psi^t , \ \mathbf{wx} \leq C \right\}$ and their functional representation via

the indirect directional output distance function (IDDF):

$$CE = ID\left(\mathbf{w}/C, \mathbf{y}; \mathbf{g}_y\right) = \sup_{\beta} \left\{\beta : \left(\mathbf{y} + \beta \mathbf{g}_y\right) \in IP\left(\mathbf{w}/C\right)\right\} \tag{4}$$

This alternative measure of inefficiency involves the use of input prices and it measures how much output production could be expanded, given that the overall cost the firm is facing is given. Since the constraint here is the overall cost rather than a specific input vector, this type of inefficiency has also been called cost constrained inefficiency (see Fare and Grosskopf, 1994; Grosskopf et al., 1997). It should be emphasized that, though the cost is constraining production, the inefficiency measure is defined on the output side for a given *numeraire* $\mathbf{g}_y$. This means that the quantities in equations (3) and (4) share the same underlying *numeraire* and they can be compared. Invoking the duality theorem proved in Fare and Primont (2006), it holds that $ID\left(\mathbf{w}/C, \mathbf{y}; \mathbf{g}_y\right) \geq D^t\left(\mathbf{x}, \mathbf{y}; \mathbf{g}_y\right)$, where the difference between the two indicators is an allocative efficiency component, interpreted as the loss in output due to the choice of a non-optimal input mix:

$$AE = ID\left(\mathbf{w}/C, \mathbf{y}; \mathbf{g}_y\right) - D^t\left(\mathbf{x}, \mathbf{y}; \mathbf{g}_y\right) \tag{5}$$

The allocative inefficiency definition embedded in equation (5) determines the quantity of output which is lost because the firm fails to choose the optimal input mix given the prevailing input prices. Given the duality between the direct and the indirect DDF, it is thus possible to decompose the firm level cost constrained inefficiency (CE) into the two components defined in equations (5) and (3):

$$CE = TE + AE \tag{6}$$

The left hand side of this equation is a measure of the overall loss in output for a firm operating at the specified cost level. The first component on the right hand side attributes part of this inefficiency to a less than optimal use of the given input vector; while the second component is a measure of loss in output attributable to the firm failing to choose an optimal

input vector.

II. Industry Technology and Inefficiency

The previous section introduced the main representation of the firm technology set and the firm inefficiency. The purpose of this section is to extend these notions to the industry level. For the purposes of this study the industry is defined in terms of inputs and outputs homogeneity, so that all the firms operating in the industry uses the same set of inputs to produce the same set of outputs. It is assumed that any number of firms can operate in the industry (entry and exit of firms is allowed) and all the firms in the industry (already operating in it or potentially entrant) face the same technology set Ψ^t defined in equation (2); in other words all the firms in the industry face the same production trade-offs. Under these assumptions the industry technology set is defined as (see Peyrache, 2013):

$$\Psi_I^t = \cup_{S=1}^{+\infty} \left(\sum_{j=1}^{S} \Psi^t \right) \tag{7}$$

The summation in parentheses is a special case of the aggregation discussed in Li and Ng (1995) and Zelenyuk (2006). It should be noted that, though the firm technology set is convex, the industry technology set may show some non-convexities. These non-convexities at the industry level arises because of the indivisibility of the firm: only an integer number of firms can operate in the industry (of course these non-convexities become less important as the number of firms grow large). The set defined in (7) can be also written as follows:

$$\Psi_I^t = \left\{ (\mathbf{x}, \mathbf{y}) : \lambda \mathbf{X}^t \leq \mathbf{x}' , \ \lambda \mathbf{Y}^t \geq \mathbf{y}' , \ \sum_k \lambda_k = S , \ \lambda \geq \mathbf{0} , \ S \in \mathbb{N} \right\} \tag{8}$$

Contrary to definition (2), the intensity vector is now constrained to sum up to the the integer number S which represents the number of firms operating (actually or potentially) in the industry. Interestingly enough, this set collapses to the firm technology set (2) when $S = 1$. The industry technology returns all the possible input and output combinations which are feasible at the industry level and it is an enlargement of the firm production set. Since the

industry as a whole is operating within a given cost budget, it is interesting to describe the industry technology set in terms of the associated indirect output sets which represent all the output combinations the industry is able to produce given the overall cost budget available. The industry indirect ouput sets are given by $IP_I(\mathbf{w}/C) = \left\{\mathbf{y} : (\mathbf{x}, \mathbf{y}) \in \Psi_I^t,\ \mathbf{wx} \leq C\right\}$ (note that this definition uses the industry technology (7) rather than the firm technology (2) as a benchmark) and their functional representation is given by the industry indirect output distance function:

$$ID_I(\mathbf{w}/C, \mathbf{y}; \mathbf{g}_y) = \sup_{\beta} \{\beta : (\mathbf{y} + \beta \mathbf{g}_y) \in IP_I(\mathbf{w}/C)\} \tag{9}$$

This new definition represents, for a given input price vector and a given cost, the overall loss in output when the benchmark is the industry technology rather than the firm technology. In other words, this definition is comparing the observed output produced to a benchmark dictated by the industry technology. Though this definition is very useful in theoretical terms, it is less than clear what prices to use when applying it empirically. Observed industry inputs, outputs and cost can be defined quite easily as the sum of inputs, outputs and costs accross the firms operating in the industry. In fact, the total output of the industry will be equal to the sum of all the firm level outputs $\left(\mathbf{Q}^t = \sum_k \mathbf{y}_k^t\right)$ and, similarly, the industry overall cost budget will be the sum of the firm level costs ($C_I = \sum_k C_k$). On the contrary (unless the law of one price applies in the inputs market) there will be a heterogeneous variety of different input price vectors which segments the inputs market and make the price vector faced by each firm different. In this general case, it is not clear how to define an industry input price vector and therefore it is not clear which input price vector to use in the definition of the industry indirect DDF (9). In order to state this problem and its solution more clearly, suppose that the inputs market is divided into J different segments and in each segment a particular input price vector prevails (this embeds the law of one price as a special case). These input price vectors are collected into an observational matrix $\mathbf{W}' = (\mathbf{w}_1', \mathbf{w}_2', \ldots, \mathbf{w}_J')$. An example of a segmented inputs market is when input prices vary accross different geographical locations within a nation or a macro region. Each firm, by deciding in which segment of the market operate production, implicitly decides what vector of input prices to face. The number

of segments of the market can be smaller, equal or larger than the number of firms actually operating in the industry. In fact, the first two cases corresponds to a situation in which every firm is facing a different input price vector or some of the firms share the same input price vector. The last case corresponds to a situation in which there are more price vectors than firms. Though it may seem an unlikely situation, this case corresponds to a counterfactual analysis in which one considers not only the price vectors that the firms are curretnly facing in the industry, but also an additional number of price vectors that will drive the counterfactual analysis. So, for example, one may be interested in assessing the potential efficiency gains obtainable by the industry via a shift of production towards locations where the industry is not currently operating. This counterfactual analysis may be important when assessing if to open a new firm (an entrant) in a segment of the inputs market where no other firms are operating yet. All in all, what is important from a mathematical point of view is to distiguish and keep separate theoretically the number of segments of the market from the number of firms which actually operate in the industry. The industry input price vector can be defined, for this general case, as a weighted average of the different segmented input prices in the following way:

$$\bar{\mathbf{w}} = \sum_{j=1}^{J} \alpha_j \mathbf{w}_j \tag{10}$$

where $\alpha_j = \frac{C_j}{\sum_j C_j}$ is the share of segment j on the total cost of the industry (note that the shares are here expressed in terms of the J segments of the market rather than the K^t firms operating in the industry). This industry price is a function of local prices $\mathbf{w}_j$ and the allocation of the total cost budget of the industry accross the different segments of the inputs market (i.e. the weights α_j). It should be noted that this industry price definition is intuitive, since it collapses to the observed price in the case where all the cost goes to one specific segment and the law of one price applies (for example, in the case $\alpha_1 = 1$, $\alpha_2 = 0, \ldots, \alpha_J = 0$ the industry price is $\bar{w} = w_1$). From an industry perspective these weights are a variable of choice, in the sense that the allocation of the industry cost budget should make profit of the differences in prices accross the different segments of the inputs market (the industry should

source the inputs from the most favourable inputs market segment). In other terms, it is possible to choose the weights in order to maximize the total potential output of the industry. This definition implicitly assumes that the number of firms that can populate each segment of the inputs market can vary; so, for example, a new entrant firm can decide in which segment to operate irrespective of how many firms are already operating in that segment. This discussion implies the following definition of industry inefficiency (IE), which takes both the number of firms and the allocation of cost accross different segments as variables of choice:

$$(11) \qquad \begin{aligned} IE^t &= ID_I^t\left(\mathbf{Q}, \mathbf{W}/C_I; \mathbf{g}_y\right) = \\ &= \sup_{\beta,\alpha}\left\{\beta : \left(\mathbf{Q}+\beta\mathbf{g}_y\right) \in IP_I\left(\mathbf{w}/C_I\right), \mathbf{w}=\alpha\mathbf{W}, \alpha \geq \mathbf{0}, \sum_{j=1}^{K}\alpha_j = 1\right\} \end{aligned}$$

The solution to this problem provides: *i)* the optimal number of firms that should populate the industry; *ii)* the optimal allocation of the total industry cost accross the different segments of the inputs market; *iii)* the overall inefficiency of the industry. The inefficiency of the industry is, once again, measured in terms of the numeraire $\mathbf{g}_y$. It should be noted that this optimization is a non-linear integer program and its full solution is provided in the appendix. The definition of industry inefficiency just provided involves two very different types of inefficiency: on one hand, the inefficiency arising from a non-optimal allocation of cost accross different firms (which includes a non-optimal number of firms as a special case); on the other hand, a failure to optimally allocate the overall cost budget accross the different segments of the inputs market. In order to disentagle these two very different effects, it is useful to provide the following restricted version of industry inefficiency (this idea of reallocation is similar to the one presented in Nesterenko and Zelenyuk, 2007):

$$(12) \qquad ISE^t = ID_I\left(\bar{\mathbf{w}}/C, \mathbf{y}; \mathbf{g}_y\right) = \sup_{\beta}\left\{\beta : \left(\mathbf{y}+\beta\mathbf{g}_y\right) \in IP_I\left(\bar{\mathbf{w}}/C\right)\right\}$$

where $\bar{\mathbf{w}}$ is defined using equation (10) and the shares are set equal to the observed ones. Since we impose a restriction on the weights α, it holds that $IE^t \geq ISE^t$, where the difference

is given by a price effect:

$$PE^t = IE^t - ISE^t$$

The price effect vanishes if the matrix of prices $\mathbf{W}$ contains only one element, i.e. in the case in which the law of one price applies. Since there is no effect of the market segmentation in definition (12), this measure returns the inefficiency of the industry which is associated to a non-optimal allocation of cost accross the different firms operating in the industry. It should be noted that, since the number of firms is a variable of choice at the industry level, the reallocation of cost embedded in (12) may well involve changing the number of firms which operate in the industry. To be more specific: if the number of firms operating in the industry is below the optimal level, the reallcoation involves reallocation of cost to new entrants; if the number of firms operating in the industry is above the optimal, reallocation involves exit of some of the firms from the industry. Of course, such a reallocation effect may still be present when the number of firms is at the optimal: this is due to the economies of scale and scope structure of the technology. Since this is an important point, it is worth stressing once again that the two components of industry inefficiency just defined correspond to two very distinct cost reallocation effects: the first one corresponds to reallocation of the industry cost budget accross firms (and potentially entrant or exiting firms); the second one is associated with a reallocation of the industry cost budget accross the different segments of the inputs market. In general the two reallocations are conceptually radically different and must be kept separate.

Industry inefficiency decomposition.

Given the previous definitions, it is now possible to seek for a decomposition of the industry inefficiency indicator defined in (11) into its sources components. The first step in this direction is to consider the impact of the inefficiencies arising from the firms already operating in the industry. These were defined in the previous section in equations (3) and (4) and can now be aggregated by a simple sum accross firms (the reader should note how simple aggregation of directional distance functions is). This is possible because an additive notion of inefficiency has been used and the *numeraire* $\mathbf{g}_y$ is common to all firms. The industry technical inefficiency is therefore defined as (see Briec et al, 2003; Fare et al, 2008; Fare and Primont, 2003;

Fare and Zelenyuk, 2003):

$$ITE^t = \sum_{k=1}^{K^t} D^t\left(\mathbf{x}_k^t, \mathbf{y}_k^t; \mathbf{g}_y\right) \tag{13}$$

Since this is only one part of the total inefficiency of the firm, one should also aggregate the cost constrained inefficiency into an industry indicator:

$$ICE^t = \sum_{k=1}^{K^t} ID^t\left(\mathbf{w}_k^t/C_k^t, \mathbf{y}_k^t; \mathbf{g}_y\right) \tag{14}$$

A decomposition of the firm inefficiency was given in the previous section and it still holds at the industry level, giving rise to an industry allocative inefficiency indicator:

$$\begin{aligned} IAE^t &= ICE^t - ITE^t = \\ &= \sum_{k=1}^{K^t} \left[ID^t\left(\mathbf{w}_k^t/C_k^t, \mathbf{y}_k^t; \mathbf{g}_y\right) - D^t\left(\mathbf{x}_k^t, \mathbf{y}_k^t; \mathbf{g}_y\right)\right] \end{aligned} \tag{15}$$

The difference between the two indicators is a measure of the impact on industry output loss of non optimal input choices at the firm level. Therefore overall industry inefficiency can be decomposed as follow:

$$IE = ISE + PE = ITE + IAE + IRE + PE \tag{16}$$

In this decomposition: the ITE component accounts for the technical inefficiency of the firms actually operating in the industry; the IAE component accounts for the overall effect of the failure to optimize the choice of input vectors by the firms actually operating in the industry; the IRE component corresponds to the effect of the potential reallocation of cost accross firms (actually or potentially) operating in the industry; and the PE component accounts for potential gains obtainable via a reallocation of the industry cost budget accross the different segments of the inputs market. A more direct interpretation of this result can

be obtained by expressing all these quantities in percentage terms:

$$\%ITE + \%IAE + \%IRE + \%PE = 1 \tag{17}$$

where $\%ITE = \frac{ITE}{IE}$, $\%IAE = \frac{IAE}{IE}$, $\%IRE = \frac{IRE}{IE}$, $\%PE = \frac{PE}{IE}$.

An important special case: the law of one price in the inputs market.

It is interesting to consider the case in which all the input prices are the same accross the different segments of the input market. This is a benchmark case because corresponds to the case in which the law of one price applies to the inputs market. In this case there is no ambiguity on which price vector to use in the definition of industry inefficiecy. The industry inefficiency definition provided in (11) therefore collapses to:

$$IE = ID_I\left(\mathbf{w}/C_I, \mathbf{Q}^t; \mathbf{g}_y\right) \tag{18}$$

where the industry price is the same as the vector of input prices faced by each firm in the industry. This definition is equivalent to the definition given in equation (12). Of course, in this special case there will be no price effect in the industry inefficiency decomposition, which consequently becomes:

$$IE = ITE + IAE + IRE \tag{19}$$

It should be noted that when one considers the program in (12) or the program of equation (18) the input prices are fixed. Since for an optimization program like this a solution always exists (under the stated conditions), the question becomes what this solution looks like in terms of industry configuration. Though uniqueness of such a solution cannot be guaranteed, following an idea presented in Maindiratta (1990) (and extended in Ray, 2007; Ray and Hu, 1997; Ray and Mukherjee, 1998), it is possible to use the convexity assumption to show that any optimal configuration of the industry is equivalent to a configuration where S firms are allocated an equal cost share and produce the same output. To prove this, suppose

that there exists a configuration of the industry where firms are allocated a different cost share $(C_1, \ldots, C_S)$, $C_i \neq C_j \; for \; some \; i, j$ and the total production is $(Q_1, \ldots, Q_S)$, $\sum Q_k = Q_{Eff}$. Then by using convexity, it is also true that it is possible to produce $\bar{Q}_{Eff} = Q_{Eff}/S$ using cost $\bar{C} = \sum C_k/S$: in fact, since the input prices are common the second expression can also be written as $\sum wx_k/S = w \sum x_k/S$ which is the average of the input vector (which is a convex linear combination of the postulated input vectors). Therefore the configuration where S firms produce the same output using the same cost is equivalent to any other possible optimal configuration. This result is quite interesting because it makes possible to express the optimal configuration in terms of averages and the industry inefficiency as:

$$ISE = S \cdot ID\left(\mathbf{w}/\bar{C}, \bar{\mathbf{Q}}\right) \tag{20}$$

This is an extension of the interpretation of industry inefficiency in terms of the average firm as firstly proposed by Forsund and Hjalmarsson (1979).

Counterfactual Analysis.

It is interesting to consider the case in which the researcher is interested in investigating a situation in which input prices change from the observed segmentation $\mathbf{W}$, to a new one say $\mathbf{W}_C$ (where the subscript stays for "counterfactual"). The question now becomes how the industry responds to such a change. Assuming that the total cost budget of the industry stays unchanged to the same level, the new optimization problem for the industry is:

$$IE_C^t = ID_I^t\left(\mathbf{Q}, \mathbf{W}_C/C_I; \mathbf{g}_y\right)$$

It should be noted that it may be the case that $IE_C^t < IE^t$ and in fact it could also be negative if counterfactual prices are set to such a level as to make overall potential production lower than the observed one. This counterfactual analysis may inform on changes associated with the elimination of subsidies for some of the production factors in the industry. An example of this counterfactual analysis is given in the empirical section.

III. Dynamics: Industry Productivity Change

The previous sections defined the static problem of finding the optimal configuration of the industry in a specific time period and then benchmarking the actual configuration to this potential in order to define an efficiency measure. Consider now the problem of determining how much the industry as a whole improves or deteriorates its performance between two time periods. A direct comparison of the industry inefficiency levels is not completely informative because the industry now faces two different technologies: the base period technology Ψ_I^t and the comparison period technology Ψ_I^{t+1}. Since these two technologies will in general differ, one shall consider the effect of technical change from one period to the next. The industry total output and total cost in the two time periods are $\left(\mathbf{Q}^t,\, \mathbf{Q}^{t+1}\right)$ and $\left(C^t,\, C^{t+1}\right)$. It is quite straightforward to define the industry inefficiency change (IEC) as the difference between the observed levels of inefficiency in the two time periods:

$$IEC = IE^t - IE^{t+1} \tag{21}$$

This indicator takes a value larger (lower) than zero if the efficiency of the industry improved (deterioreted) and the industry is operating closer (farther) to the optimal configuration. Naturally, the optimal configuration benchmark in the two time periods is different and this must be accounted for when considering if the industry improved (or deterioreted) its performance. A measure of technical change is given by the shift of the production frontier measured at the base period or comparison period output-cost point. The base period industry technical change (ITC) measure is given by:

$$ITC^{t+1} = ID_I^{t+1}\left(\mathbf{Q}^t, \mathbf{W}^t/C_I^t; \mathbf{g}_y\right) - ID_I^t\left(\mathbf{Q}^t, \mathbf{W}^t/C_I^t; \mathbf{g}_y\right) \tag{22}$$

Interestingly enough, this definition relies not only on observed cost and output quantities, but also on the structure of the base period input price vector. In fact, a different segmentation of the inputs market will return a different measure of technical change, by implicitly changing the point at which the displacement of the production frontier is measured. An alternative

technical change measure is given by considering the comparison period:

$$ITC^t = ID_I^{t+1}\left(\mathbf{Q}^{t+1}, \mathbf{W}^{t+1}/C_I^{t+1}; \mathbf{g}_y\right) - ID_I^t\left(\mathbf{Q}^{t+1}, \mathbf{W}^{t+1}/C_I^{t+1}; \mathbf{g}_y\right) \tag{23}$$

A similar argument applies here, with the technical change component depending on the structure of the segmented inputs market. The technical change indicators take a value larger (smaller) than zero when the technology frontier shifts outwards (inwards) signalling technical progress (regress). Since there are two alternative indicators of technical change, one may follow the tradition of the Malmquist-Luenberger literature by taking the average of the previous two indexes as a measure of technical change. Productivity change is then nothing else than the sum of technical change and inefficiency change at the industry level:

$$IProd = ITC + IEC \tag{24}$$

This indicator will be larger (smaller) than zero if the industry experienced an improvement (decline) in productivity. Since the previous sections provided a decomposition of industry inefficiency into sources components, it is possible to insert this decomposition into equation (24) in order to obtain a more fine decomposition of industry productivity change:

$$IProd = ITC + ITEC + IAEC + IREC + PEC \tag{25}$$

The components of this decomposition are defined as: $ITEC = ITE^t - ITE^{t+1}$, $IAEC = IAE^t - IAE^{t+1}$, $IREC = IRE^t - IRE^{t+1}$, $PEC = PE^t - PE^{t+1}$.

IV. Empirical Illustration

In this section the methodology will be applied to data on 92 electricity distributors in Ontario, Canada for T=7 years from 2005 to 2011. These data provide information on 3 outputs and 3 inputs. The outputs are: electricity distributed, other products and services and transmission losses. The inputs are labour (in full time equivalent), capital services and other materials services (intermediate inputs). Information on the labour input is only

available for 2011, which complicates things a little bit. In fact, the omission of a relevant input would cause the estimated inefficiency to be higher than it should be (see Fare and Zelenyuk, 2002). Though this is an obvious problem, it is still useful to use the data to illustrate the methodology, keeping in mind that the dynamic results about productivity change are affected by this omission. It is worth mentioning the inclusion of transmission losses on the output side. This variable highly correlates with the level of electricity distributed and the transmission losses in the data are of the order of between 1% to 10% of the total electricity distributed. Prices for inputs were obtained from different sources. Monthly wage rates and producer and consumer price indexes were obtained from Statistics Canada. Bond yields on capital (which is a measure of the cost of capital) were obtained from the Bank of Canada. The remaining data were extracted from the Yearbooks of Electricity Distributors published by the Ontario Energy Board (OEB). All price and quantity measures combine to yield the total revenue, total expenses and earnings before interest and taxes reported in the Ontario Energy Board (OEB) yearbooks. Figure A1 reports boxplots of the variables used standardized by the sample mean. It is quite clear from this picture that the Ontario electricity distribution system show quite a high degree of heterogeneity on the size of firms, with a few big firms and a large number of smaller firms. Since the main output produced in the industry is the electricity distributed, this is reflected in the choice of the directional vector: $\mathbf{g}_y = \begin{pmatrix} 1 & 0 & 0 \end{pmatrix}$. This means that the projection involves an expansion of the quantity of electricity distributed towards the frontier, while keeping the other outputs constant. This also means that the unit of measurement for the inefficiency indicators discussed in this section is the same as the unit of measurement of the electricity distributed.

Table A1 reports the results for the industry inefficiency decomposition (16) in each year. Since Labour input is only available for 2011, the decomposition is reported both with the inclusion and the exclusion of the labour input for this year; while in all other years the labour input is omitted (because it is missing in the dataset). Table A2 reports the same results in percentage terms following equation (17). The total inefficiency of the industry accross the different years is quite stable. Using the last line of table A1 the total inefficiency of the system is equal to 1.93E+11 KwH of loss in electricity for the given overall cost budget of the

industry. The meaning of this figure is that with the observed overall budget of the industry, if all the inefficiencies were to be eliminated, that very cost would be able to dispacth almost two times the quantity of energy that it is currently distributing. In fact by taking the ratio of actual to potential production gives an inefficiency score of around 0.4, which means the industry is operating well below its potential. One may look at table A1 for the contribution of the different components to this overall inefficiency. But a much more direct interpretation can be obtained from table A2. In this table it is quite clear that in 2011 (when labour is included) 79.3% of this overall inefficiency is given by a reallocation effect, which implies a transfer of cost accross the different firms in the industry in order to optimally exploit scale and scope economies and bring the number of firms operating in the industry to the optimal. The remaining 20.7% of the industry inefficiency is due to the inefficiency of the firms actually operating in the industry. According to equation (6) there are two components that induce this type of inefficiency. The technical inefficiency component (ITE) represents lack of best practices, while the allocative inefficiency component (IAE) represents a failure to choose optimal input quantities given the input prices. It is clear from these figures that the IAE component is the most important component in the overall firm inefficiency. These results change a little bit if one looks at years were the labour input was not available, though the change is not dramatic and the main story stays unchanged. Summing up, the main component in the inefficiency of the industry is a reallocation effect and the second most important component is a failure to optimize with respect to the prevalent input prices.

Table A3 provides the dynamic results associated with the decomposition of the industry productivity change as explained in equation (25). The productivity of the industry as a whole has been quite erratic, jumping forward and backward from one year to the next. It should be noted though that the order of magnitude of this change has been of 1 or 2 times smaller than the overall output of the industry or the overall inefficiency of the industry. In other words, though there were jumps up and down, these jumps have been small with respect to the overall production of the industry. The table reports also the different components which underlie this productivity change. The first component is the techncal change component (TC) and it represents how much the frontier has been shifting outwards (or inwards). From

the numbers reported in the table it is clear that the frontier has been shifting outwards in the first 3 years and then it has been going inwards for two years.

A more interesting result is the one reported in table A4. Here a counterfactual analysis is conducted based on the following idea. It is well known that there is a structure of subsidies and regulations in the Ontario electricity distribution sector which is influencing the choices of the agents operating in it. Without the ambition of doing any justice to the complexity of the system, here two aneddotical facts are used to run a counterfactual analysis. The first has to do with regulation on the cost of capital, with regulation imposing some type of cap on the return on equity. It is well known that this type of regulation provides an incentive for manager to expand the capital stock above the optimal level in order to maintain the correct proportion of return on equity. In other words this type of regulation implies, de facto, a reduction in the price of capital. The other aneddotical stylized fact is that worker unions in the energy sector were able to bargain higher salaries with respect to the average of other sectors, due to the special structure of subsidies of this sector. This implicitly result in a higher than average price of labour in the electricity distribution sector. Based on this aneddotical evidence the counterfactual analysis assumes a scenario in which wages decline by 10% and the cost of capital increases. Table A4 reports the results for such an exercise. The table reports input prices and the associated optimal input quantities. This optimal input quantities are determined as the best response to the prevailing input prices and are the quantities that should be chosen by the average firm (20) in order to make the industry operate at its potential. The table reports how these optimal quantities vary when the input prices are varied from the observed ones to the counterfactual ones (and the cost is kept fixed at the observed one). The capital input is the one that shows the highest percentage reduction (6.5%) followed by labour (6.1%) and intermediate materials (5.5%). This reduction in optimal input quantities induce a reduction in the overall potential output of the industry of 6.7%. It is also interesting to note that the optimal cost shares of the various inputs change dramatically with the capital share increasing from 14.6% to 22.1% and the share of the other two inputs declining. It should be noted that the reduction in the output of the industry is a reduction in the potential output. In fact, since the observed output is well below this

optimal, the reduction induced by the counterfacutal could be offset by an increase in the inudsty efficiency. Therefore a central planner that would want to untight regulation in the sector could do so, according to this counterfactual analysis, by providing at the same time stronger incentives towards efficiency, so that the reduction in the subsidies would not affect the overall output of the industry.

V. Conclusion

In this paper a definition of industry inefficiency in cost constrained production environments is introduced. This definition makes use of the Indirect Directional Distance Fucntion (IDDF) and quantifies the inefficiency of the industry in terms of overall output loss for a given overall industry cost budget. The methodology proposed in this paper identifies the optimal configuartion of the inudstry and shows how to decompose it into sources components: inefficiencies arising from sub-optimal configuration of the industry (reallocation inefficiency); inefficiency arising from a failure to optimize input quantities given the prevalent inputs prices; inefficiencies due to lack of best practices at the firm level. These notions are then extended to the dynamic case in order to define the industry productivity change and its associated decomposition. The method is illustrated using data on Ontario electricity distributors. The data shows that lack of best practices is only a minor component of the overall inefficiency of the sector (less than 10%), with reallocation inefficiency accounting for more than 75% of the overall inefficiency of the system. A counterfactual analysis based on counterfactual inputs prices is conducted in order to illustrate how the model can be used to estimate the effects of a change in the regulation regime.

REFERENCES

Banker R.D., Chang H., Natarajan R. (2005), "Productivity change, technical progress and relative efficiency change in the public accounting industry", *Management Science*, 51, 291-304.

Banker R.D., Charnes A., Cooper W.W. (1984), "Some models for estimating technical and scale inefficiencies in data envelopment analysis", *Management Science*, 30, 1078-1092.

Briec W., Dervaux B., Leleu H. (2003), "Aggregation of directional distance functions and industrial efficiency", *Journal of Economics*, 79, 237-261.

Chambers R. G., Chung Y., Fare R. (1996), "Benefit and distance functions", *Journal of Economic Theory*, 70, 407-419.

Chambers R.G., Chung Y., Fare R. (1998), "Profit, directional distance functions and Nerlovian efficiency", *Journal of Optimization Theory and Applications*, 98, 351-364.

Fare R., Grosskopf S. (1994), "Cost and revenue constrained production", Springer-Verlag.

Fare R., Grosskopf S., Zelenyuk V. (2008), "Aggregation of Nerlovian profit indicator", *Applied Economics Letters*, 15, 845-847.

Fare R., Primont D. (2003), "Luenberger productivity indicators: aggregation across firms", *Journal of Productivity Analysis*, 20, 425-435.

Fare R., Primont D. (2006), "Directional duality theory", *Economic Theory*, 29, 239-247.

Fare R., Zelenyuk V. (2002), "Input aggregation and technical efficiency", *Applied Economics Letters*, 9, 635-636.

Fare R., Zelenyuk V. (2003), "On aggregate Farrell efficiencies", *European Journal of Operational Research*, 146, 615-620.

Farrell M.J. (1957), "The measurement of productive efficiency", *Journal of the Royal Statistical Society*, 120, 253-290.

Forsund F.R., Hjalmarsson L. (1979), "Generalised Farrell measures of efficiency: an application to milk processing in Swedish dairy plants", *The Economic Journal*, 89, 294-315.

Grosskopf S., Hayes K.J., Taylor L.L., Weber W.L. (1997), "Budget-constrained frontier

measures of fiscal equality and efficiency in schooling", *The Review of Economics and Statistics*, 79, 116-124.

Li S.K., Ng Y.C. (1995), "Measuring the productive efficiency of a group of firms", *International Advances in Economic Research*, 1, 377-390.

Nesterenko V., Zelenyuk V. (2007), "Measuring potential gains from reallocation of resources", *Journal of Productivity Analysis*, 28, 107-116.

Maindiratta A. (1990), "Largest size-efficient scale and size efficiencies of decision-making units in data envelopment analysis", *Journal of Econometrics*, 46, 57-72.

Peyrache A. (2013), "Industry structural inefficiency and potential gains from mergers and break-ups: a Comprehensive approach", *European Journal of Operational Research*, 230, 422-430.

Ray S.C. (2007), "Are some Indian banks too large? An examination of size efficiency in Indian banking", *Journal of Productivity Analysis*, 27, 41-56

Ray S.C., Hu X. (1997), "On the technically efficient organization of an industry: a study of U.S. airlines", *Journal of Productivity Analysis*, 8, 5-18.

Ray S.C., Mukherjee K. (1998), "Quantity, quality and efficiency for a partially superadditive cost function: Connecticut public schools revisited", *Journal of Productivity Analysis*, 10, 47-62.

Zelenyuk V. (2006), "Aggregation of Malmquist productivity indexes", *European Journal of Operational Research*, 174, 1076-1086.

Tables and Figures

Table A1—Industry Inefficiency Decomposition.

Year	IE	ISE	IRE	ICE	IAE	ITE
2005	2.24E+11	2.24E+11	1.87E+11	3.67E+10	1.90E+10	1.77E+10
2006	2.45E+11	2.45E+11	2.23E+11	2.25E+10	1.39E+10	8.55E+09
2007	2.63E+11	2.63E+11	2.38E+11	2.57E+10	1.71E+10	8.57E+09
2008	2.59E+11	2.59E+11	2.33E+11	2.50E+10	1.44E+10	1.06E+10
2009	2.25E+11	2.25E+11	2.04E+11	2.12E+10	3.06E+09	1.82E+10
2010	1.99E+11	1.99E+11	1.80E+11	1.91E+10	9.64E+09	9.46E+09
2011	2.43E+11	2.43E+11	2.07E+11	3.60E+10	2.37E+10	1.23E+10
2011(Labour)	1.93E+11	1.93E+11	1.53E+11	4.00E+10	2.80E+10	1.20E+10

Table A2—Industry Inefficiency Decomposition (Percentages).

Year	IRE	ICE	IAE	ITE
2005	83.6%	16.4%	8.5%	7.9%
2006	90.8%	9.2%	5.7%	3.5%
2007	90.3%	9.7%	6.5%	3.3%
2008	90.3%	9.7%	5.6%	4.1%
2009	90.6%	9.4%	1.4%	8.1%
2010	90.4%	9.6%	4.8%	4.7%
2011	85.2%	14.8%	9.7%	5.1%
2011(Labour)	79.3%	20.7%	14.5%	6.2%

Table A3—Industry Productivity Change Decomposition.

Year	IProd	TC	IEC	ITEC	IAEC	IREC
2006	2.50E+10	4.66E+10	-2.16E+10	9.17E+09	5.06E+09	-3.58E+10
2007	-7.03E+09	1.12E+10	-1.82E+10	-1.50E+07	-3.16E+09	-1.50E+10
2008	6.08E+10	5.59E+10	4.93E+09	-2.07E+09	2.73E+09	4.27E+09
2009	-1.66E+11	-1.99E+11	3.31E+10	-7.52E+09	1.13E+10	2.93E+10
2010	-2.63E+10	-5.25E+10	2.62E+10	8.70E+09	-6.58E+09	2.41E+10
2011	1.23E+11	1.67E+11	-4.38E+10	-2.83E+09	-1.40E+10	-2.70E+10

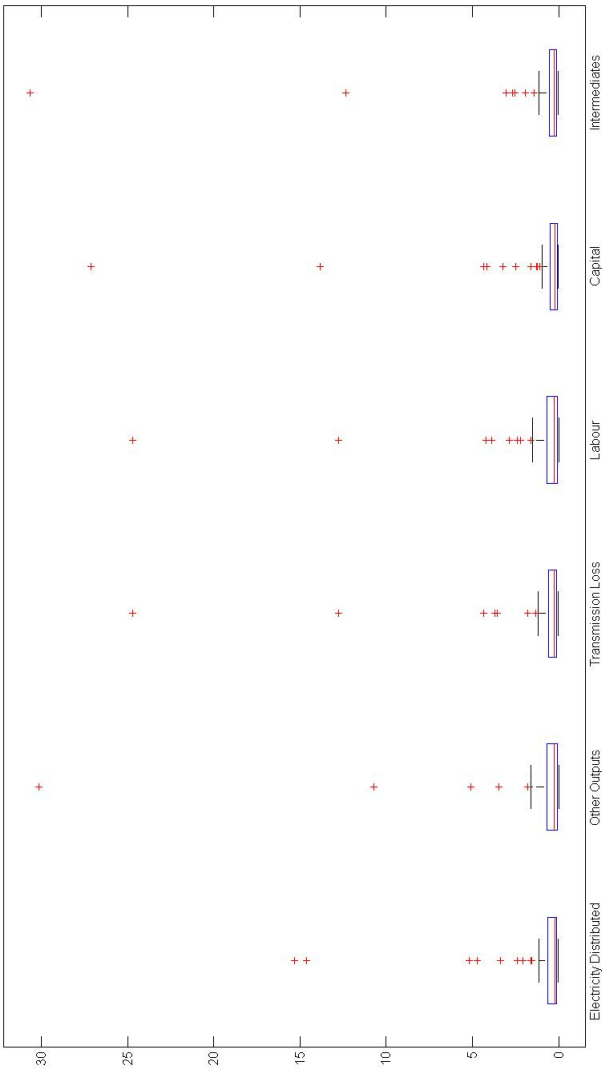

Figure A1. Descriptive Boxplot; Variables standardized by sample means (Year 2011).

Table A4—Industry Inefficiency Counterfactual.

	Labour	Capital	Intermediates
Input prices (Observed)	38,500	3.09	178
Input prices (Counterfactual)	34,700	5.00	178
Input Quantities (Optimal for Observed)	1.64E+04	1.06E+08	7.16E+06
Input Quantities (Optimal for Counterfactual)	1.54E+04	9.86E+07	6.76E+06
Percentage reduction in Input quantities	6.1%	6.5%	5.5%
Input Cost shares (Optimal for Observed)	28.3%	14.6%	57.1%
Input Cost shares (Optimal for Counterfactual)	23.9%	22.1%	53.9%

Total Industry Output	Inefficiency (Observed)	Inefficiency (Counterfactual)	Potential Output reduction (Percentage)
1.21E+11	1.93E+11	1.7221E+11	6.7%

APPENDIX: Optimization programs

The purpose of this appendix is to provide the optimization programs associated with the definitions used in the text in a more explicit way. The firm direct DDF (3) is obtained solving the following linear program:

$$
\begin{aligned}
D\left(\mathbf{x},\mathbf{y};\mathbf{g}_y\right) = sup_{\beta,\lambda,\mathbf{x}} \quad & \beta \\
s.t.: \quad & \lambda\mathbf{X}^t \leq \mathbf{x}' \\
& \lambda\mathbf{Y}^t \geq \mathbf{y}' + \beta\mathbf{g}_y' \\
& \sum_k \lambda_k = 1 \\
& \lambda \geq \mathbf{0}
\end{aligned}
\tag{B1}
$$

The firm indirect DDF defined in (4) corresponds to the solution of a linear program, which can be explicitly written as follow:

$$
\begin{aligned}
ID\left(\mathbf{w}/C,\mathbf{y};\mathbf{g}_y\right) = sup_{\beta,\lambda,\mathbf{x}} \quad & \beta \\
s.t.: \quad & \lambda\mathbf{X}^t \leq \mathbf{x}' \\
& \lambda\mathbf{Y}^t \geq \mathbf{y}' + \beta\mathbf{g}_y' \\
& \sum_k \lambda_k = 1 \\
& \lambda \geq \mathbf{0} \\
& \mathbf{wx} \leq C
\end{aligned}
\tag{B2}
$$

The ISE component of equation (12) can be computed solving the following integer linear

program:

$$
\begin{aligned}
ISE = sup_{\lambda,S,\mathbf{x}} \quad & \beta \\
s.t.: \quad & \lambda \mathbf{X}^t \leq \mathbf{x}' \\
& \lambda \mathbf{Y}^t \geq \mathbf{Q}' + \beta \mathbf{g}_y' \\
& \sum_k \lambda_k = S \\
& S \in \mathbb{N} \\
& \lambda \geq \mathbf{0} \\
& \bar{\mathbf{w}} \mathbf{x} \leq C
\end{aligned}
\tag{B3}
$$

The input price vector is the weighted average defined in (10) with the shares fixed to the observed ones and $\mathbf{Q}$ is the total observed industry output. To solve this integer linear program, it is sufficient to solve its associated CRS specification:

$$
\begin{aligned}
sup_{\lambda,\mathbf{x}} \quad & \beta \\
s.t.: \quad & \lambda \mathbf{X}^t \leq \mathbf{x}' \\
& \lambda \mathbf{Y}^t \geq \mathbf{Q}' + \beta \mathbf{g}_y' \\
& \lambda \geq \mathbf{0} \\
& \bar{\mathbf{w}} \mathbf{x} \leq C
\end{aligned}
\tag{B4}
$$

This will give a value for the intensity vector $\sum_k \lambda_k$. The full problem can then be solved for the nearest integer numbers close to the value of the intensity vector provided by the CRS program; the proper solution will be the one which returns the highest value for the objective function.

The industry inefficiency indicator contained in equation (11) corresponds to the following

non-linear integer program:

$$
\begin{aligned}
IE = sup_{\alpha,S,\lambda,\mathbf{x}} \quad & \beta \\
s.t.: \quad & \lambda \mathbf{X}^t \leq \mathbf{x}' \\
& \lambda \mathbf{Y}^t \geq \mathbf{Q}' + \beta \mathbf{g}_y \\
& \sum_k \lambda_k = S \\
& \lambda \geq \mathbf{0} \\
& \left(\sum_j \alpha_j \mathbf{w}_j \right) \mathbf{x} \leq C \\
& 0 \leq \alpha_j \leq 1 \\
& \sum_j \alpha_j = 1
\end{aligned}
\tag{B5}
$$

where $\mathbf{w}_j$ are the elements of matrix $\mathbf{W}$ and S is fixed to the optimal value determined previously. All the other components of the industry inefficiency decomposition can be determined with the information provided by this optimization programs.