A generalized inefficiency model with input and output dependence

Mike G. Tsionas*

May 10, 2023

Abstract

In this paper we propose a general inefficiency model, in the sense that technical inefficiency is, simultaneously, a function of all inputs, outputs, and contextual variables. We recognize that change in inefficiency is endogenous or rational, and we propose an adjustment costs model with firm-specific but unknown adjustment cost parameters. When inefficiency depends on inputs and outputs, the firm's optimization problem changes as the first order conditions must take into account the dependence of inefficiency on the endogenous variables of the problem. The new formulation introduces statistical challenges which are successfully resolved. The model is estimated using Maximum Simulated Likelihood and an empirical application to U.S. banking is provided.

Key Words: Productivity and Competitiveness; Adjustment Costs; Statistical Endogeneity; Maximum Simulated Likelihood.

Acknowledgments: The author is grateful to two anonymous reviewers for useful comments on an earlier version.

*Montpellier Business School & Lancaster University Management School, LA1 4YX, U.K., m.tsionas@montpellier-bs.com

1 Introduction

In this paper we consider technical inefficiency as a function of inputs and outputs. The dominant paradigm in the literature is to use environmental or contextual variables as determinants of technical inefficiency (Battese and Coelli, 1988, 1995, Battese and Broca, 1997, and Wang, 2002). Such determinants are useful when one wishes to argue that inefficiency may depend on the context, regulation, etc. However, it is known that achieving higher efficiency levels is costly in terms of resources and, therefore, in terms of inputs and / or outputs (Bogetoft and Hougaard, 2003). Using inputs and outputs as determinants of inefficiency opens up new problems and is by no means trivial as in the case of incorporating environmental or contextual variables. The reason is that most behavioral assumptions (cost minimization, profit or revenue maximization) require that inputs and / or outputs are selected optimally: When inefficiency depends on inputs and / or outputs, the decision making unit (DMU) optimization problem changes, as the first order conditions must take into account the dependence of inefficiency on the endogenous variables of the optimization problem.

Our paper is related to Hampf (2017), who used sequential definitions of the production technology, to decompose cost inefficiency into rational and residual inefficiency as well as inefficiency caused by technical change. Related literature includes Fukuyama and Matousek (2018), Aparicio, Mahlberg, Pastor, and Sahoo (2014), Kapelko and Oude Lansink (2017), Kapelko, Oude Lansink, and Stefanou (2014), Tran and Tsionas (2016), Tsionas and Mamatzakis (2019), etc. In addition, Hampf (2017) provided lower bounds for unobserved adjustment costs based on unexploited cost reductions due to rational inefficiency. Specifically, based on the decomposition into technical and allocative rational inefficiency, lower bounds for the radial and the non-radial adjustment costs can be estimated. Hampf (2017) suggested that parametric models could also be used, as in Park and Lesourd (2000). This study is in this direction of research, although we provide a completely new parametric model.

2 Preliminaries

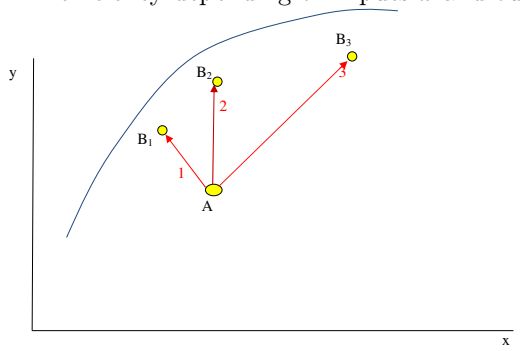
Suppose $x \in \Re_+^K$ is a vector of inputs, and we have a production function of the form:

$$y = F(x)e^{v-u}, \tag{1}$$

where $u \geq 0$ represents technical inefficiency and v is statistical noise. To describe inefficiency in a general input - output space we need to assume $u = u(x, y)$. To see this consider Figure 1. From a point A it is possible to move to points B_1, B_2 or B_3. Point B_2 corresponds to a vertical movement (direction 2) corresponding to no change in the inputs. This change is rarely possible and points like B_1 or B_3 are more likely to occur in practice. This, however, requires simultaneous changes in both inputs and outputs.

In addition, to move from A to any other point like B_3 requires real resources in terms of both inputs and outputs, viz. it is costly to adjust inputs and outputs. A move from point A to B_1 (direction 1) could occur when adjustment costs are high and inputs need to be reduced: In this case, however, output can be increased (direction 1) when, for example, managerial

Figure 1: Inefficiency depending on inputs x and outputs y



practices improve. This increase in output may be also subject to adjustment costs when management (and, therefore, its cost) is unobserved. These adjustment costs are used to evaluate the feasibility of exploiting cost reductions caused by inefficiency.

Point A could be interpreted as last period's input-output combination and, in turn, moving to points like B_1, B_2 or B_3, amounts to selecting a particular movement in the input-output space. Such movements (with the exception of "2") are, however, costly and, in practice, the costs are unobserved. Moreover, it is not necessary to assume that such movements result in efficiency improvements and, in general, it is clear that inefficiency should depend on both inputs and outputs. Suppose that we write:

$$\ln y = f(\ln x) + v - u(\ln x, \ln y), \tag{2}$$

where $f(\ln x) = \ln F(x)$. We should notice that the distribution of y conditional on x depends also on y through $u(x, y)$. This may seem to prevent further analysis of the model as y appears on both sides of (2). Technically, what is required to obtain the distribution of y conditional on x is the Jacobian of transformation:

$$\frac{\partial v}{\partial \ln y} = 1 + \frac{\partial u(\ln x, \ln y)}{\partial \ln y}, \tag{3}$$

so that the distribution has density:

$$p_{\ln y}(\ln y) = p_v\left(\ln y - f(\ln x) + u(\ln x, \ln y)\right) \cdot |1 + \frac{\partial u(\ln x, \ln y)}{\partial \ln y}|. \tag{4}$$

Inefficiency dependence on both inputs and outputs can be motivated based on Figure 1 where inefficiency changes require changes in both inputs and outputs. These changes are, however, costly, in real life since increasing or decreasing inputs (and the associated changes in outputs) is, generally speaking, costly. A typical example is when labor or capital have to be increased or decreased: Increases take time and are costly while decreases may be subject to labor laws or disinvestment which is also time-consuming as well as costly. In particular, less utilization of equipment services implies costs of depreciation among others like maintenance and repair and similar costs are in place when equipment services have to be increased. Labor service utilization can be increased only with time and obvious additional costs per unit. These considerations imply that, in general, input-output changes are costly and resemble a situation close to the one in Figure 1. An alternative is to use a directional distance function,

see Chambers, Chung, and Färe (1996), and Chung, Färe, and Grosskopf (1997). Deriving optimal directions has received some attention, see Peyrache and Daraio (2012) who recommend sensitivity analysis to different direction vectors, and Färe, Grosskopf, and Whittaker (2013) who propose to estimate the optimal directions by maximizing inefficiency with respect to the direction vectors. This line of research has been explored in Atkinson, Primont, and Tsionas (2018). The methodology presented here is different in several respects. First, we can estimate directly adjustment costs. Second, directions are implicitly defined as the movement from point A to points like B_1, B_2 or B_3 can be estimated. Third, these directions are endogenized in the model via the adjustment costs for each input and output. Fourth, technical inefficiency is directly a function of inputs, outputs and control or contextual variables, so that effects of input-output changes on inefficiency can be estimated directly. This introduces statistical challenges which are successfully resolved. To the best of our knowledge, there are no models that can incorporate the influence of input-output-contextual variables on technical inefficiency, simultaneously.

3 The model

Suppose $x \in \Re_+^K$ is a vector of inputs, $y \in \Re_+^M$ is a vector of outputs and $z \in \Re^J$ is a vector of control or environmental variables. Feasible combinations of inputs and outputs are described by:

$$\mathcal{T} = \left\{ x \in \Re_+^K, y \in \Re_+^M, z \in \Re^J | x, y, z \text{ can be produced} \right\}. \tag{5}$$

The firm maximizes profits:

$$\max_{(x,y,z)\in\mathcal{T}} : p'y - w'x, \tag{6}$$

where $w \in \Re_+^K$ is a vector of input prices and $p \in \Re_+^M$ and is a vector of output prices. We represent the technology using an output distance function (ODF, Fare and Primont, 1995, p. 11):

$$D(x, y, z) = \min \{\vartheta : (x, y/\vartheta, z) \in \mathcal{T}\} \tag{7}$$

It is non-decreasing, positively linearly homogeneous, increasing and convex in y, and decreasing and quasi-concave in x. Moreover, $D(x, y) \leq 1 \, \forall (x, y) \in \mathcal{T}$ with $D(x, y) = 1$ if and only if (x, y) is efficient. If TE is the Farrell-type output oriented measure of technical efficiency, then

$$D(x, y) = 1/TE. \tag{8}$$

Alternatively, we can write:

$$D(x, y) = e^{-u} \leq 1, \tag{9}$$

where $u \geq 0$ represents technical inefficiency. In this paper we assume that technical inefficiency may be a function of inputs, outputs and the control variables:

$$u = u(x, y, z) \, \forall (x, y, z) \in \mathcal{T}. \tag{10}$$

The problem of the firm instead of (6) is:

$$\begin{gathered}\max_{(x,y,z)} : p'y - w'x - \tfrac{1}{2}\sum_{m=1}^{M} c_m^y p_m (y_m - y_m^o)^2 - \tfrac{1}{2}\sum_{k=1}^{K} c_k^x w_k (x_k - x_k^o)^2 - \tfrac{1}{2}\sum_{j=1}^{J} c_j^z (z_j - z_j^o)^2 \\ \text{subject to } D(x,y,z) = e^{-u(x,y,z)}.\end{gathered} \tag{11}$$

The adjustment cost coefficients (for outputs and inputs) are expressed in terms of prices of outputs and inputs. In (11), y^o, x^o and z^o are the previous period values of outputs, inputs and the controls. There are adjustment costs c_m^y for each output, c_x^x for each input, and c_j^z for each control. If a variable can be adjusted without cost we can simply set the respective adjustment cost to zero. If it is beyond the control of the form we can set it to infinity or a very large number. The first order conditions to the problem are as follows:

$$\begin{aligned} p_m \{1 - c_m^y (y_m - y_m^o)\} &= \lambda \left\{ \tfrac{\partial D}{\partial y_m} + \tfrac{\partial u}{\partial y_m} e^{-u} \right\} \forall m = 1, ..., M, \\ w_k \{1 + c_k^x (x_k - x_k^o)\} &= -\lambda \left\{ \tfrac{\partial D}{\partial x_k} + \tfrac{\partial u}{\partial x_k} e^{-u} \right\} \forall k = 1, ..., K, \\ c_j^z (z_j - z_j^o) &= -\lambda \left\{ \tfrac{\partial D}{\partial z_j} + \tfrac{\partial u}{\partial z_j} e^{-u} \right\} \forall j = 1, ..., J. \end{aligned} \tag{12}$$

We can express these conditions in equivalent form as follows:

$$\begin{aligned} p_m y_m \{1 - c_m^y (y_m - y_m^o)\} &= \lambda e^{-u} \left\{ \tfrac{\partial \ln D}{\partial \ln y_m} + \tfrac{\partial \ln u}{\partial \ln y_m} u \right\} \forall m = 1, ..., M, \\ w_k x_k \{1 + c_k^x (x_k - x_k^o)\} &= -\lambda e^{-u} \left\{ \tfrac{\partial \ln D}{\partial \ln x_k} + \tfrac{\partial \ln u}{\partial \ln x_k} u \right\} \forall k = 1, ..., K, \\ c_j^z z_j (z_j - z_j^o) &= -\lambda e^{-u} \left\{ \tfrac{\partial \ln D}{\partial \ln z_j} + \tfrac{\partial \ln u}{\partial \ln z_j} u \right\} \forall j = 1, ..., J, \end{aligned} \tag{13}$$

using (9). Denote $\frac{\partial \ln D}{\partial \ln y_m} = \varepsilon_{y_m,}^D$, $\frac{\partial \ln u}{\partial \ln y_m} = \varepsilon_{y_m}^u$, etc. In turn, we have:

$$\begin{aligned} p_m y_m \{1 - c_m^y (y_m - y_m^o)\} &= \lambda e^{-u} \left\{ \varepsilon_{y_m}^D + \varepsilon_{y_m}^u u \right\} \forall m = 1, ..., M, \\ w_k x_k \{1 + c_k^x (x_k - x_k^o)\} &= -\lambda e^{-u} \left\{ \varepsilon_{x_k}^D + \varepsilon_{x_k}^u u \right\} \forall k = 1, ..., K, \\ c_j^z z_j (z_j - z_j^o) &= -\lambda e^{-u} \left\{ \varepsilon_{z_j}^D + \varepsilon_{z_j}^u u \right\} \forall j = 1, ..., J. \end{aligned} \tag{14}$$

To eliminate the Lagrange multiplier λ we have:

$$\begin{aligned} \tfrac{\tilde{p}_m y_m \{1 - c_m^y (y_m - y_m^o)\}}{y_1 \{1 - c_1^y (y_1 - y_1^o\}} &= \tfrac{\varepsilon_{y_m}^D + \varepsilon_{y_m}^u u}{\varepsilon_{y_1}^D + \varepsilon_{y_1}^u u} \forall m = 2, ..., M, \\ \tfrac{\tilde{w}_k x_k \{1 + c_k^x (x_k - x_k^o)\}}{x_1 \{1 + c_1^x (x_1 - x_1^o)\}} &= \tfrac{\varepsilon_{x_k}^D + \varepsilon_{x_k}^u u}{\varepsilon_{x_1}^D + \varepsilon_{x_1}^u u} \forall k = 2, ..., K, \\ \tfrac{\tilde{c}_j^z z_j (z_j - z_j^o)}{z_1 (z_1 - z_1^o)} &= \tfrac{\varepsilon_{z_j}^D + \varepsilon_{z_j}^u u}{\varepsilon_{z_1}^D + \varepsilon_{z_1}^u u} \forall j = 2, ..., J, \end{aligned} \tag{15}$$

where $\tilde{p}_m = p_m / p_1$, $\tilde{w}_k = w_k / w_1$, and $\tilde{c}_j^z = c_j^z / c_1^z$. This is a system of $M + K + J - 3$ equations in $M + K + J$ endogenous

variables. The two of the three missing equations are provided by

$$\begin{array}{c}\lambda = \frac{p_1 y_1\{1-c_1^y(y_1-y_1^o)\}e^u}{\left\{\varepsilon_{y_1}^D+\varepsilon_{y_1}^u u\right\}} = \\ -\frac{w_1 x_1\{1+c_1^x(x_1-x_1^o)\}e^u}{\varepsilon_{x_1}^D+\varepsilon_{x_1}^u u} = \\ -\frac{c_1^z z_1(z_1-z_1^o)e^u}{\varepsilon_{z_1}^D+\varepsilon_{z_1}^u u},\end{array} \tag{16}$$

that is we have:

$$\begin{array}{c}\frac{y_1\{1-c_1^y(y_1-y_1^o)\}}{\varepsilon_{y_1}^D+\varepsilon_{y_1}^u u} = -\frac{x_1\{1+c_1^x(x_1-x_1^o)\}}{\varepsilon_{x_1}^D+\varepsilon_{x_1}^u u}, \\ \frac{y_1\{1-c_1^y(y_1-y_1^o)\}}{\varepsilon_{y_1}^D+\varepsilon_{y_1}^u u} = -\frac{z_1(z_1-z_1^o)}{\varepsilon_{z_1}^D+\varepsilon_{z_1}^u u},\end{array} \tag{17}$$

where we set $w_1 = c_1^z = 1$ without loss of generality, for normalization.

The last equation is given by (9). The system consisting of (9), (15) and (17) can be written in the form:

$$\mathbf{F}(Y; c, u, \theta) = 0, \tag{18}$$

where $Y = [y', x', z']'$, $\mathbf{F}$ is a vector function in $\Re^{M+K+J}$, c is the vector of adjustment costs, and $\theta \in \Theta \subseteq \Re^d$ is the vector of parameters in the ODF. If we have panel data, the adjustment coefficients are firm specific, and technical inefficiency is firm-specific and time-varying we have:

$$\mathbf{F}(Y_{it}; c_i, u_{it}, \theta) = v_{it}\ \forall i = 1, ..., n, t = 1, ..., T, \tag{19}$$

where

$$v_{it}|Y_{it} \sim \mathcal{N}_{M+K+J}(\mathbf{0}, \Sigma). \tag{20}$$

The first element of Σ is σ_v^2 and other elements in the first row and column of Σ are zero. Before proceeding we need to parametrize technical inefficiency and the ODF. For the ODF we assume a translog functional form. This can be written as:

$$\ln Y_{it,1} = \beta_0 + \sum_{j=1}^{N-1} \beta_j \ln \tilde{Y}_{it,j} + \sum_{j=1}^{N-1}\sum_{j'=1}^{N-1} \beta_{jj'} \ln \tilde{Y}_{it,j} \ln \tilde{Y}_{it,j'} + v_{it} - u_{it}, \tag{21}$$

where $\tilde{Y}_{it} = [y_{it,2}/y_{it,1}, ..., y_{it,M}/y_{it,1}, x_{it,1}, ..., x_{it,K}, z_{it,1}, ..., z_{it,J}] \in \Re^{N-1}$, $N = M + J + K$.[1] For technical inefficiency we assume:

$$u_{it} \sim \mathcal{N}_+(\mu_{it}, \sigma_u^2), \tag{22}$$

i.e., it follows a truncated normal distribution, where μ_{it} is also a translog:

$$\mu_{it} = \gamma_0 + \sum_{j=1}^{N-1} \gamma_j \ln \tilde{Y}_{it,j} + \sum_{j=1}^{N-1}\sum_{j'=1}^{N-1} \gamma_{jj'} \ln \tilde{Y}_{it,j} \ln \tilde{Y}_{it,j'}. \tag{23}$$

[1] Alternatively, we have estimated a Generalized Leontief functional form but the results are omitted as they were quite close to those of the translog. Given the empirical data, this suggests some robustness to the functional form.

Estimation relies on Maximum Simulated Likelihood (MSL) whose details are presented in the Technical Appendix. Quantities of interest like returns to scale, technical efficiency, technical efficiency change and productivity growth can be estimated as usual (see notes to Figure 4). Given the truncated normal specification in (22), we can define inefficiency as: $RI_{it} = \mathbb{E}(u_{it})$, where $\mathbb{E}$ denotes (conditional) expectation. From Kumbhakar and Lovell (2000, p. 86) we have: $\mathbb{E}(u_{it}) = \sigma_* \left\{ \frac{\tilde{\mu}_{it}}{\sigma_*} + \frac{\varphi(\tilde{\mu}_{it}/\sigma_*)}{\Phi(\tilde{\mu}_{it}/\sigma_*)} \right\}$, where $\varphi(\cdot)$ is the standard normal density function, $\sigma_*^2 = \frac{\sigma_v^2 \sigma_u^2}{\sigma_v^2 + \sigma_u^2}$, $\tilde{\mu}_{it} = \frac{\mu_{it}\sigma_v^2 - \sigma_u^2 \varepsilon_{it}}{\sigma_v^2 + \sigma_u^2}$, $\varepsilon_{it} = v_{it} - u_{it}$ in (21) (at the estimated parameters), and σ_v^2 is the uppermost left element of Σ.

4 Data and empirical results

The data is the same as in Malikov, Kumbhakar, and Tsionas (2016). We use data on commercial banks from Call Reports available from the Federal Reserve Bank of Chicago and include all FDIC insured commercial banks with reported data for 2001:I-2010:IV. We focus on a selected subsample of relatively homogeneous large banks, viz. those with total assets in excess of one billion dollars (in 2005 U.S. dollars) in the first three years of observation. The data sample is an unbalanced panel with 2,397 bank-year observations for 285 banks. We have the following desirable outputs: consumer loans (y_1), real estate loans (y_2), commercial and industrial loans (y_3) and securities (y_4). These output categories are the same as those in Berger and Mester (1997, 2003). Following Hughes and Mester (1998, 2013), we include off-balance-sheet income (y_5) as output. The undesirable output is total non-performing loans (NPL). The variable inputs are labor, i.e., the number of full-time equivalent employees (x_1), physical capital (x_2), purchased funds (x_3), interest-bearing transaction accounts (x_4) and non-transaction accounts (x_5). We also include financial (equity) capital (EQ) as an additional input to the production technology. However, due to the unavailability of the price of equity capital, we follow Berger and Mester (1997, 2003) and Feng and Serletis (2009) in modeling EQ as quasi-fixed. This is in line with Hughes and Mester's (1993, 1998) in that banks may use equity as a source of funds. We derive the prices of variable inputs (w_1 through w_5) by dividing total expenses on each input by the respective input quantity.

In Figure 2 we report sample distributions of adjustment cost parameters and inefficiency effects. Adjustment cost parameters are lowest for labor (x_1), followed by purchased funds (x_3), interest-bearing transaction accounts (x_4) and physical capital (x_2) which has the largest adjustment coefficient, which makes intuitive sense on prior grounds as equipment and capital, in general, cannot be adjusted easily. The reason why this makes sense is because the model captures at least some part of what is going on in the real world with input and output adjustments. This is, conventionally, ignored and interest focuses on inefficiencies without crediting the firm with rational (profit-maximizing) adjustments in the input-output space. However, ignoring these adjustments overestimates (irrational) inefficiency and does not allow for inefficiency reductions realized through changes in outputs and inputs.

Inefficiency effects in the form of elasticities ($\frac{\partial \log E(u)}{\partial \log Y}$) are, for the most part, positive for all outputs and negative for all inputs with the exception of x_2. Output adjustment costs are highest for NPLs and lowest for y_4 (securities) with other adjustment costs lying somewhere in between. For selected inputs and outputs, sample distributions by percentile are shown in Figure 3. It is important to mention that these elasticities serve as bank-specific directional parameters in the input-output space as in Figure

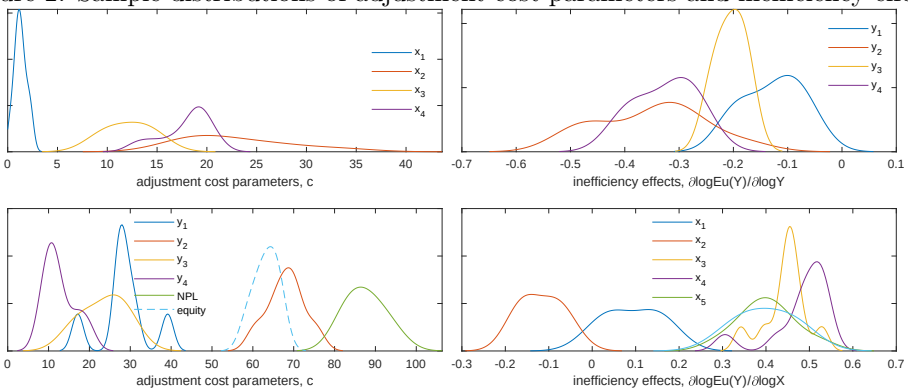
Table 1: Adjustment costs

	Value	s.d.
		$10^{-2}\times$
	Outputs	
y_1	2.75%	1.23
y_2	3.44%	0.87
y_3	2.81%	1.33
y_4	4.77%	1.45
y_5	5.22%	0.28
	Inputs	
x_1	3.73%	0.67
x_2	9.45%	0.86
x_3	5.82%	0.71
x_4	7.82%	1.55
NPL	8.77%	2.33

1. For the most part, and based also on the results reported in Figure 3, the directions are positive for inputs and negative for outputs suggesting an adjustment process (relative to Figure 1) that is more complicated relative to the straightforward northwest or northeast directions. With more than one input and output this adjustment process is necessarily more complex. The results of Figure 3 can be useful when bank-specific directions are sought on a quantile-based basis. To understand the nature of adjustment cost parameters better we convert the c^y and c^x into percentages in terms of the corresponding input and output values. The results are reported in Table 1. From these results we see that output adjustment costs are 4-5% but input adjustment costs are more variable ranging from 3% to just under 10%. The adjustment costs due to NPLs average to 8.77% (s.d. 2.33%). Since we have 285 banks with total assets more than a billion USD, these adjustment costs are sizable, giving in turn a justification for inefficiency which is about 15% on the average. These real costs in other words, imply that relative movements in the input-output space are costly and inefficiency cannot be reduced at will. In a certain sense this is why the presence of inefficiency is "rational" and is justified by the technology or cost / profit structure of US commercial banks. From the policy point of view it seems relatively unlikely that inefficiency and adjustment costs are due to the "quiet life hypothesis" which states that commercial banks can trade some inefficiency for the monopolist's "quiet life". Inefficiency appears to be due to the impossibility of certain adjustments in the input-output space or inertia due to the technology and profit-maximizing behavior. This is why an argument like "reduce inefficiency first and bail out later" may not work for commercial banks as the amount of reductions in inefficiency is rather limited by the technology and profit-maximizing behavior. Surely, *some* amount of inefficiency can be reduced directly but *for the most part* inefficiency is just the way things are given the technology and behavioral assumptions about commercial banks in the US.

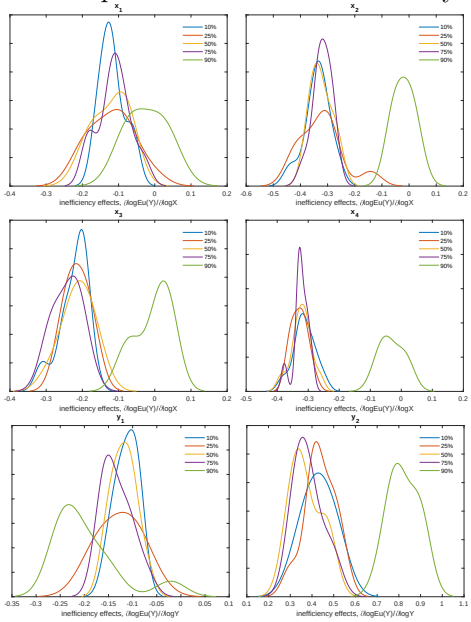
The distribution of weights w_s at the MSL parameter estimates $\hat{\theta}$ is reported in Figure A1 of Technical Appendix (left panel). In the right panel of Figure A1 we report 20 representative (kernel) densities of the weights corresponding to specific observations (i,t) across all simulations at the MSL parameter estimates $\hat{\theta}$. In Figure 4, reported is the sample distribution of technical inefficiency (upper left), technical change (upper right), efficiency change (lower left) and productivity growth (lower right). Cost efficiency is $r_{it} = e^{-u_{it}}$ where u_{it} is technical inefficiency. Technical change is measured by $e_{y,it}$ which is the elasticity

Figure 2: Sample distributions of adjustment cost parameters and inefficiency effects



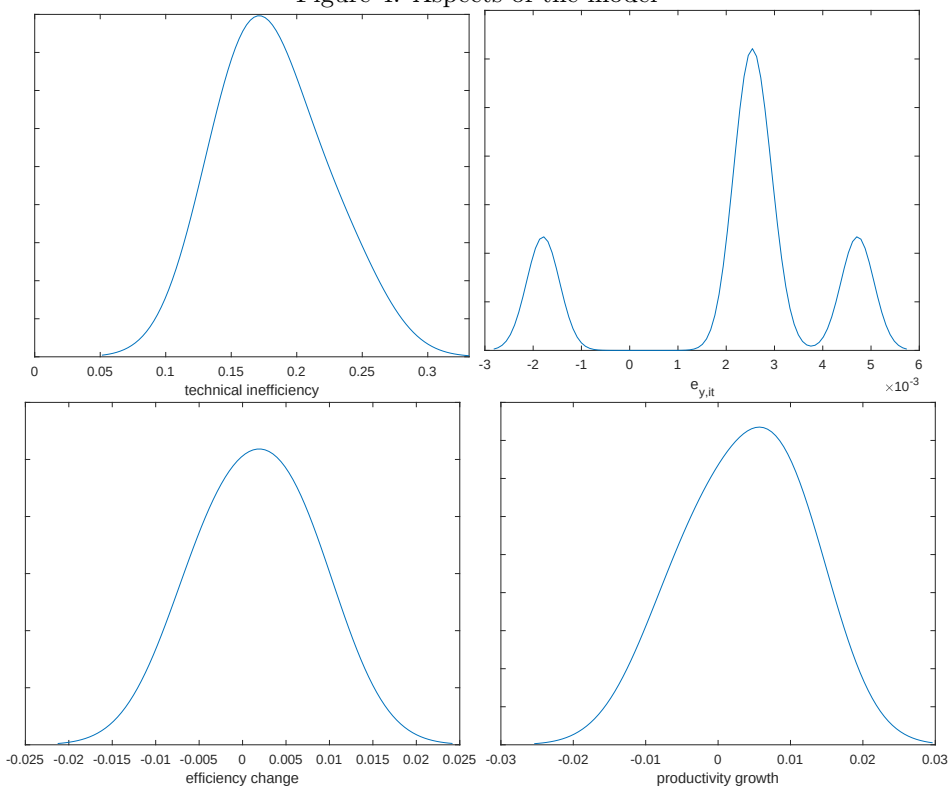
Notes: All adjustment cost parameters are relative to x_5 (non-transaction accounts). We have the following desirable outputs: consumer loans (y_1), real estate loans (y_2), commercial and industrial loans (y_3) and securities (y_4). These output categories are the same as those in Berger and Mester (1997, 2003). Following Hughes and Mester (1998, 2013), we include off-balance-sheet income (y_5) as output. The undesirable output is total non-performing loans (NPL). The variable inputs are labor, i.e., the number of full-time equivalent employees (x_1), physical capital (x_2), purchased funds (x_3), interest-bearing transaction accounts (x_4) and non-transaction accounts (x_5). We also include financial (equity) capital (e) as an additional input to the production technology.

Figure 3: Sample distributions of inefficiency effects



Notes: All adjustment cost parameters are relative to x_5 (non-transaction accounts). We have the following desirable outputs: consumer loans (y_1), real estate loans (y_2), commercial and industrial loans (y_3) and securities (y_4). These output categories are the same as those in Berger and Mester (1997, 2003). Following Hughes and Mester (1998, 2013), we include off-balance-sheet income (y_5) as output. The undesirable output is total non-performing loans (NPL). The variable inputs are labor, i.e., the number of full-time equivalent employees (x_1), physical capital (x_2), purchased funds (x_3), interest-bearing transaction accounts (x_4) and non-transaction accounts (x_5). We also include financial (equity) capital (e) as an additional input to the production technology.

Figure 4: Aspects of the model



Notes: Cost efficiency is $r_{it} = e^{-u_{it}}$ where u_{it} is technical inefficiency. Technical change is measured by $e_{y,it}$ which is the elasticity of (21) with respect to time. Efficiency change is $EC_{it} = \ln(r_{it}/r_{i,t-1})$. Productivity growth is $PG_{it} = e_{y,it} + EC_{it}$.

of (21) with respect to time. Efficiency change is $EC_{it} = \ln(r_{it}/r_{i,t-1})$. Productivity growth is $PG_{it} = e_{y,it} + EC_{it}$. Efficiency change is positive for some banks and negative for others, and the same is true for productivity growth. Moreover, technical inefficiency ranges from 5% to 35%.

5 Concluding remarks

In this paper we propose a generalized (distance function) model where inefficiency depends on both inputs and outputs. Statistical challenges are involved as, to put things in a simplified matter, in a production function model, inefficiency depends on both inputs and outputs. The challenges from dependence on output are successfully resolved. Moreover, we recognize that inefficiency can be rational due to adjustment costs involved in the re-structuring in the input-output space. The model is successfully applied to a data set of large U.S. banks. Limitations of the model, and thus possible avenues for future research include, among others, conversion of parts of the model in semiparametric or nonparametric components. The semiparametric component may lift the parametric specification in either the form of the ODF, the inefficiency specification part or possibly both. This involves identification problems and may not be easy in practice. A fruitful avenue for future research may also involve more general adjustment cost processes, possibly in a nonparametric way, which would enrich our understanding of inefficiency and associated changes in the input-space of the firm. From the policy viewpoint, and to keep in mind for future applications, our inefficiency estimates take into account and, therefore, they are robust to rational, profit-maximizing, changes

in inputs and outputs. It would be interesting to come up with new measures of rational inefficiency where overall inefficiency is decomposed into rational and residual (or "irrational"). One possible way of doing this is to follow Hampf (2017) and define residual inefficiency as deviation from its conditional expectation. Although this is certainly possible it would be much more interesting to build residual inefficiency directly into the model.

Technical Appendix

The density corresponding to (22) is:

$$p(u_{it}) = (2\pi)^{-1/2}\sigma_u^{-1}\Phi(-\mu_{it}/\sigma_u)^{-1}\exp\left(-\frac{(u_{it}-\mu_{it})^2}{2\sigma_u^2}\right),\ u_{it}\geq 0, \tag{A.1}$$

where $\Phi(\cdot)$ is the standard normal distribution function. Define the Jacobian of transformation:

$$\mathcal{J}(Y_{it};c_i,u_{it},\theta) = ||\partial\mathbf{F}(Y_{it};c_i,u_{it},\theta)/\partial Y_{it}||. \tag{A.2}$$

Due to the complexity of (19) the Jacobian is computed numerically. Define $c=[c_1',...,c_n']'$ and $u=[u_{it},i=1,...,n,t=1,...,T]$. Then the likelihood function is:

$$\mathscr{L}(\theta,c,\Sigma;Y)\propto|\Sigma|^{-nT/2}\int_{\Re_+^{nT}}\exp\left\{-\tfrac{1}{2}\sum_{i=1}^{n}\sum_{t=1}^{T}\mathbf{F}(Y_{it};c_i,u_{it},\theta)'\Sigma^{-1}\mathbf{F}(Y_{it};c_i,u_{it},\theta)\right\}\prod_{i=1}^{n}\prod_{t=1}^{T}\mathcal{J}(Y_{it};c_i,u_{it},\theta)p(u_{it};\theta)du. \tag{A.3}$$

where $p(u_{it};\theta)$ is the density of technical inefficiency in (22). Concentrating with respect to Σ we have:

$$\mathscr{L}(\theta,c;Y)\propto\int_{\Re_+^{nT}}||\mathbf{A}(Y;c,u,\theta)||^{-nT/2}\prod_{i=1}^{n}\prod_{t=1}^{T}\mathcal{J}(Y_{it};c_i,u_{it},\theta)p(u_{it};\theta)du, \tag{A.4}$$

where $\mathbf{A}(Y;c,u,\theta)=\sum_{i=1}^{n}\sum_{t=1}^{T}\mathbf{F}(Y_{it};c_i,u_{it},\theta)\mathbf{F}(Y_{it};c_i,u_{it},\theta)'$. To estimate the model we use Maximum Simulated Likelihood (MSL). Specifically, given an importance density function $\mathcal{I}(u_{it};\alpha)$ where $\alpha\in\mathcal{A}\subseteq\Re^{d_\alpha}$, if $\left\{u_{it}^{(s)},s=1,...,S\right\}$ is a set of S i.i.d draws from the distribution whose density is $I(u_{it};\alpha)$, we can approximate (A.4) using:

$$\mathscr{L}(\theta,c;Y)\propto S^{-1}\sum_{s=1}^{S}||\mathbf{A}(Y;c,u^{(s)},\theta)||^{-nT/2}\prod_{i=1}^{n}\prod_{t=1}^{T}\mathcal{J}(Y_{it};c_i,u_{it}^{(s)},\theta)\frac{p(u_{it}^{(s)};\theta)}{\mathcal{I}(u_{it}^{(s)};\alpha)}, \tag{A.5}$$

As importance density we choose a product of truncated normal densities, $\mathcal{N}_+(\tilde{\mu}_{it},\tilde{\sigma}_u^2)$. To determine $\tilde{\mu}_{it}$ and $\tilde{\sigma}_u^2$ we assume:

$$\tilde{\mu}_{it}\sim\mathcal{N}(\tilde{a}_1,\tilde{b}_1^2),\ \ln\sigma_u^2\sim\mathcal{N}(\tilde{a}_2,\tilde{b}_2^2). \tag{A.6}$$

Then the problem reduces to finding a good setting for the parameters $\alpha=\left[\tilde{a}_1,\tilde{a}_2,\tilde{b}_1,\tilde{b}_2\right]$. Our strategy is to draw $\tilde{\mu}_{it}$

and σ_u^2 randomly from (A.6), compute the importance weights:

$$w_{it}^{(s)} = \frac{p(u_{it}^{(s)};\theta)/\mathcal{I}(u_{it}^{(s)};\alpha)}{\sum_{s'=1}^{S} p(u_{it}^{(s')};\theta)/\mathcal{I}(u_{it}^{(s')};\alpha)}, \tag{A.7}$$

and examine whether $\left\{w_{it}^{(s)}, s = 1, ..., S\right\}$ is not degenerate in the sense that most weights are zero and only a few are close to unity. In practice, we consider $w_s = (nT)^{-1} \sum_{i=1}^{n} \sum_{t=1}^{T} w_{it}^{(s)}$ and, in turn, examining whether at least 30% of these weights are non-zero. As these weights depend on θ we first choose α so that the weights w_s are not degenerate for the initial values of θ. Next we examine them for degeneracy at the final estimates. If they are degenerate we proceed with another draw for α and we repeat the same process until non-degeneracy is obtained. After a number of iterations of this procedure final weights were non-degenerate and we satisfied ourselves that their distribution was not too far away from a uniform, using visual means. The final distribution could be described well by a *beta* distribution with parameters 3.42 and 1.75, approximately. The simulated likelihood is optimized using a subspace-search algorithm.[2]Initial conditions are obtained from a version of the model without adjustment costs and without dependence of technical inefficiency on x, y and z. Finally, the function:

$$f(x, y, z) = D(x, y, z) - \mathbb{E}e^{-u(x,y,z)}, \tag{A.8}$$

must satisfy the standard theoretical properties of ODF (non-decreasing, positively linearly homogeneous, increasing and convex in y, and decreasing and quasi-concave in x) along with $f(x, y, z) \leq 0 \, \forall (x, y, z) \in \mathcal{T}$. Since $D(x, y, z)$ is in levels but $Ee^{-u(x,y,z)}$ has a different functional form, the separate components of the function in (A.8), in principle, can be identified. Without the expectation, both $D(x, y, z)$ and $e^{-u(x,y,z)}$ are quadratic functions in x, y, z and no separate identification is possible if both are translog (or another common functional form for that matter). To examine identification in practice, we consider minus the Hessian of the log likelihood:

$$\mathcal{H} = -\frac{\partial^2 \ln \mathscr{L}(\theta, c; Y)}{\partial(\theta, c)\partial(\theta, c)'}, \tag{A.9}$$

a positive-definite matrix near the maximum of the log likelihood function. The determinant of the Hessian should be non-zero if we have identification, so, we can consider the smallest eigenvalue of $\mathcal{H}$ (say λ_{min}) which should be bounded away from zero. As this depends on the data Y, we average across all observations.

[2]Subplex is a subspace-searching simplex method for the unconstrained optimization of general multivariate functions. Like the Nelder-Mead simplex method it generalizes, the subplex method is well suited for optimizing noisy objective functions. The number of function evaluations required for convergence typically increases only linearly with the problem size, so for most applications the subplex method is much more efficient than the simplex method. We use the `Fortran 77` implementation in `netlib`.

Figure A.1: Distribution of importance weights from Maximum Simulated Likelihood
Left panel: The distribution of weights w_s at the MSL parameter estimates $\hat{\theta}$. Right panel: 20 representative kernel densities of the weights corresponding to specific observations (i, t) across all simulations at the MSL parameter estimates $\hat{\theta}$.

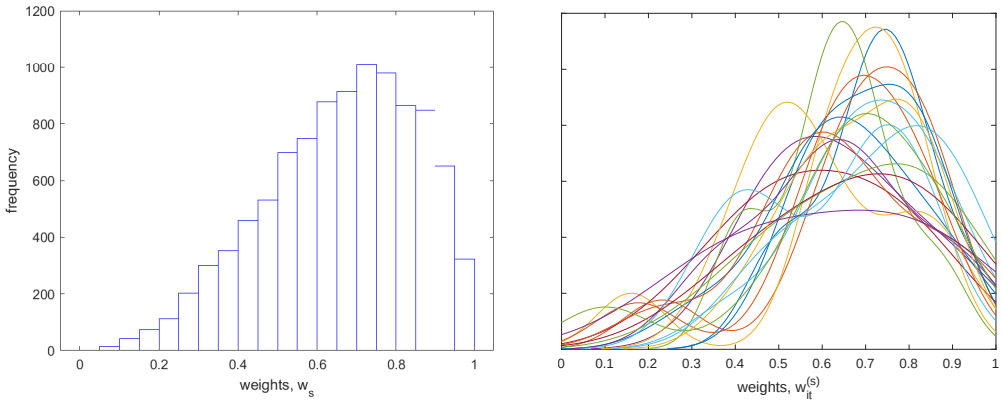
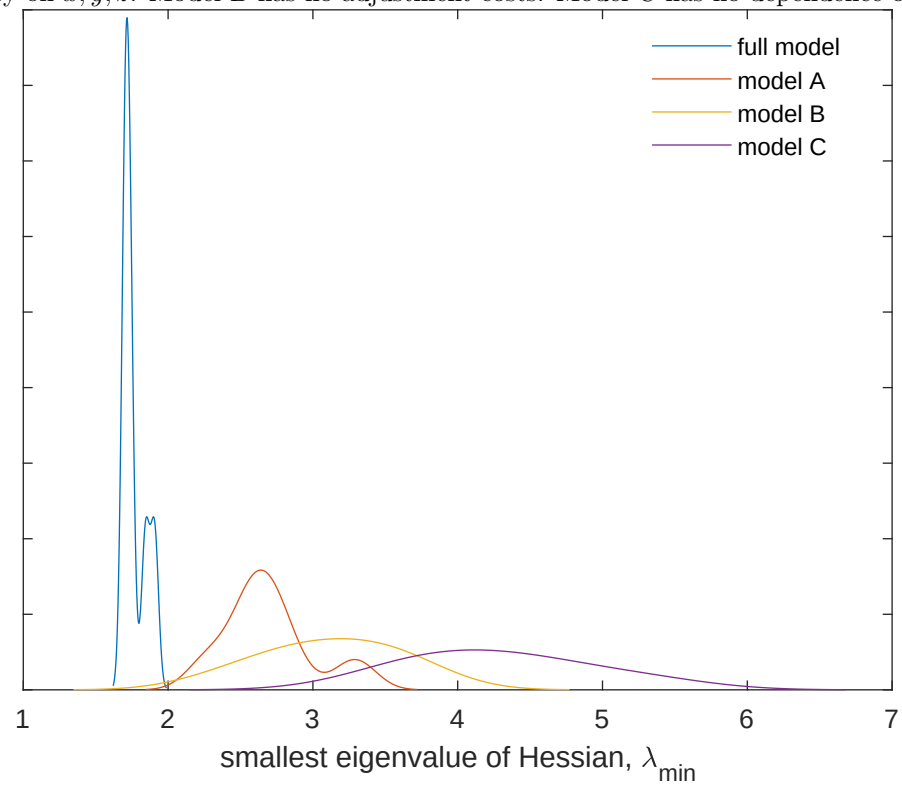


Figure A.2: Identification (sample distribution of λ_{min})
Notes: Model A has no dependence of inefficiency on x, y, z. Model B has no adjustment costs. Model C has no dependence of inefficiency on x, y, z and no adjustment costs.



Notes: To examine identification in practice, we consider minus the Hessian of the log likelihood: $\mathcal{H} = -\frac{\partial^2 \ln \mathscr{L}(\theta,c;Y)}{\partial(\theta,c)\partial(\theta,c)'}$, a positive definite matrix near the maximum of the log likelihood function. The determinant of the Hessian should be non-zero if we have identification so we can consider the smallest eigenvalue of $\mathcal{H}$ (say λ_{min}) which should be bounded away from zero. As this depends on the data Y, we can average across all observations.

Finally, we report parameter estimates of the model in Table A1 for the ODF parameters and Table A2 for the frontier parameters..

References

Aparicio, J., Mahlberg, B., Pastor, J. T., & Sahoo, B. K. (2014). Decomposing technical inefficiency using the principle of least action. European Journal of Operational Research, 239 (3), 776–785.

Atkinson, S., Primont, D., and M. G. Tsionas (2018). Statistical inference in efficient production with bad inputs and outputs using latent prices and optimal directions. Journal of Econometrics, 204 (2), 131–146.

Battese, G.E. and T.J. Coelli (1988) Prediction of Firm-Level Technical Efficiencies with a Generalized Frontier Production Function and Panel Data, Journal of Econometrics 38: 387-99.

Battese, G.E. and T.J. Coelli (1995) A Model of Technical Inefficiency Effects in a Stochastic Production Function for Panel Data, Empirical Economics 20: 325- 332.

Battese, G.E. and S.S. Broca (1997) Functional Forms of Stochastic Frontier Production Functions and Models for Technical Inefficiency Effects: A Comparative Study for Wheat Farmers in Pakistan, Journal of Productivity Analysis 8: 395-414.

Berger, A. N. and Mester, L. J. (1997). Inside the black box: What explains differences in the efficiencies of financial institutions? Journal of Banking & Finance, 21 (7), 895–947.

Berger, A. N. and Mester, L. J. (2003). Explaining the dramatic changes in performance of US banks: Technological change, deregulation, and dynamic changes in competition. Journal of Financial Intermediation, 12 (1), 57–95.

Bogetoft, P. , & Hougaard, J. L. (2003). Rational inefficiencies. Journal of Productivity Analysis, 20, 243–271 .

Casu B. and Molyneux, P. (2003). A comparative study of efficiency in European banking. Applied Economics 35 (17), 1865–1876.

Chambers R.G., Chung Y., Färe R. (1996). Benefit and Distance Functions, Journal of Economic Theory, 70, 407–19.

Chung Y.H., Färe R., Grosskopf S. (1997) Productivity and Undesirable Outputs: A Directional Distance Function Approach, Journal of Environmental Management, 51, 229–240.

Färe, R., Primont, D., 1995. Multi-output Production and Duality: Theory and Applications. Kluwer Academic Publishers, Boston.

Färe R., Grosskopf S., Whittaker G. (2013). Directional Output Distance Functions: Endogenous Constraints Based on Exogenous Normalization Constraints, Journal of Productivity Analysis, 40, 267–269.

Feng, G. and Serletis, A. (2014). Undesirable outputs and a primal Divisia productivity index based on the directional output distance function. Journal of Econometrics.

Fukuyama, H., & Matousek, R. (2018) Nerlovian revenue inefficiency in a bank production context: Evidence from Shinkin banks. European Journal of Operational Research, 271 (1), 317–330.

Hampf, B. (2017). Rational inefficiency, adjustment costs and sequential technologies. European Journal of Operational Research, 263 (3), 1095–1108.

Hughes, J. P. and Mester, L. J. (1993). A quality and risk-adjusted cost function for banks: Evidence on the \too-big-to-fail doctrine. Journal of Productivity Analysis, 4(3): 293–315.

Hughes, J. P. and Mester, L. J. (1998). Bank capitalization and cost: Evidence of scale economies in risk management and signaling. Review of Economics and Statistics, 80(2): 314–325.

Kapelko, M., & Oude Lansink, A. (2017). Dynamic multi-directional inefficiency analysis of European dairy manufacturing firms European Journal of Operational Research, 257 (1), 338–344.

Kapelko, M., Oude Lansink, A., & Stefanou, S. E. (2014). Assessing dynamic inefficiency of the Spanish construction sector pre- and post-financial crisis. European Journal of Operational Research, 237 (1), 349–357.

Koutsomanoli-Filippaki, A. & E. Mamatzakis (2009). Performance and Merton-type default risk of listed banks in the EU: A panel VAR approach, Journal of Banking & Finance 33 (11), 2050–2061.

Malikov, E., Kumbhakar, S. C., and Tsionas, M. G. (2016). A Cost System Approach to the Stochastic Directional Technology Distance Function with Undesirable Outputs: The Case of U.S. Banks in 2001-2010, Journal of Applied Econometrics, 31 (7), 1407–1429.

Park, S.-U. , & Lesourd, J.-B. (2000). The efficiency of conventional fuel power plants in south korea: A comparsion of parametric and non-parametric approaches. International Journal of Production Economics, 63, 59–67 .

Peyrache A., Daraio C. (2012). Empirical Tools to Assess the Sensitivity of Directional Distance Function to Direction Selection, Applied Economics, 44, 933–943.

Sealey, C. and Lindley, J. (1977). Inputs, outputs and a theory of production and cost of depository financial institutions. Journal of Finance 32, 1251–266.

Tran, K. C., & Tsionas, M. G. (2016). Zero-inefficiency stochastic frontier models with varying mixing proportion: A semiparametric approach European Journal of Operational Research, 249 (3), 1113–1123.

Tsionas, M. G., & Mamatzakis, E. (2019). Further results on estimating inefficiency effects in stochastic frontier models European Journal of Operational Research, 275 (3), 1157–1164.

Wang, H.J. (2002) Heteroscedasticity and Non-Monotonic Efficiency Effects of a Stochastic Frontier Model, Journal of Productivity Analysis 18: 241-253.

Table A1. Translog ODF parameters

Reported are estimates β_j and β_{jj}, of the translog functional form. The first 10 are the linear terms and the remaining are interactive terms.

	estimate	s.e.
1	-0.0423	0.0055
2	- 0.1774	0.0034
3	- 0.0073	0.0030
4	-0.1195	0.0134
5	0.0123	0.0039
6	0.0031	0.0094
7	0.2430	0.0103
8	0.0056	0.0025
9	0.0652	0.0069
10	0.1725	0.0019
11	-0.0895	0.0009
12	-0.1228	0.0067
13	0.1673	0.0083
14	0.1211	0.0090
15	-0.1338	0.0013
16	0.0493	0.0035
17	-0.2904	0.0078
18	-0.2644	0.0022
19	-0.0773	0.0037
20	-0.0317	0.0174
21	0.1352	0.0118
22	0.0883	0.0082
23	-0.0304	0.0022
24	-0.0366	0.0177

25	-0.1838	0.0033
26	-0.0750	0.0061
27	0.0891	0.0061
28	0.0300	0.0084
29	-0.1878	0.0018
30	-0.0023	0.0021
31	-0.0346	0.0028
32	0.1003	0.0014
33	0.1086	0.0075
34	0.0135	0.0029
35	0.0162	0.0040
36	0.1331	0.0111
37	0.1820	0.0177
38	-0.0491	0.0095
39	-0.0736	0.0027
40	0.0807	0.0075
41	-0.0495	0.0051
42	-0.0269	0.0077
43	-0.1285	0.0027
44	-0.0191	0.0028
45	-0.0246	0.0135
46	0.1085	0.0040
47	0.2692	0.0056
48	0.0871	0.0083
49	0.0275	0.0052
50	0.0136	0.0078
51	-0.0559	0.0013
52	-0.0746	0.0173
53	-0.0401	0.0025

54	0.0860	0.0077
55	-0.3195	0.0009
56	-0.1174	0.0080
57	0.1149	0.0124
58	-0.0411	0.0003
59	0.1136	0.0001
60	0.0588	0.0049
61	0.1335	0.0023
62	-0.0391	0.0020
63	-0.0374	0.0116
64	0.0558	0.0222
65	0.0318	0.0096

Table A2. Frontier parameters

Reported are estimates γ_j and γ_{jj}, of the translog frontier functional form. The first 10 are the linear terms and the remaining are interactive terms.

	estimate	s.e.
1	-0.2106	0.0062
2	-0.0158	0.0185
3	-0.0539	0.0024
4	-0.0980	0.0156
5	0.0411	0.0014
6	0.1196	0.0024
7	0.0508	0.0020
8	0.0296	0.0033
9	0.1400	0.0034
10	0.0345	0.0063
11	0.0787	0.0008
12	0.0198	0.0101
13	-0.0381	0.0054
14	0.0510	0.0159
15	0.0009	0.0033
16	-0.0834	0.0036
17	-0.1854	0.0090
18	0.1152	0.0060
19	0.0483	0.0113
20	0.1212	0.0026
21	-0.0188	0.0064
22	-0.1790	0.0039
23	0.0057	0.0033
24	0.0653	0.0204

25	0.0335	0.0111
26	-0.1803	0.0110
27	-0.1008	0.0004
28	0.0580	0.0001
29	-0.0334	0.0104
30	0.0353	0.0039
31	0.1192	0.0053
32	-0.0189	0.0019
33	-0.0923	0.0083
34	0.3020	0.0016
35	0.0926	0.0040
36	-0.0196	0.0011
37	-0.0079	0.0119
38	0.0806	0.0055
39	-0.0691	0.0179
40	-0.1917	0.0084
41	0.0412	0.0163
42	-0.0571	0.0081
43	0.0398	0.0023
44	0.0989	0.0153
45	-0.0153	0.0047
46	-0.1220	0.0046
47	-0.1271	0.0042
48	0.1082	0.0014
49	-0.1640	0.0120
50	0.1593	0.0073
51	0.0446	0.0082
52	0.1559	0.0034
53	0.0731	0.0029

54	-0.1014	0.0063
55	-0.1197	0.0104
56	0.1342	0.0093
57	-0.1001	0.0017
58	0.0559	0.0147
59	-0.1140	0.0128
60	-0.0383	0.0001
61	0.0906	0.0140
62	-0.0637	0.0002
63	0.0505	0.0122
64	0.0602	0.0118
65	-0.0435	0.0050