A Semiparametric Stochastic Input Distance Frontier Model with Application to the Indonesian Banking Industry*

Kai Sun[1] and Ruhul Salim[†2]

[1]School of Economics, Shanghai University, Shanghai, China
[2]School of Economics and Finance, Curtin University, Australia

Abstract

This paper proposes a semiparametric smooth-varying coefficient input distance frontier model with multiple outputs and multiple inputs, panel data, and determinants of technical inefficiency for the Indonesian banking industry during the period 2000 to 2015. The technology parameters are unknown functions of a set of environmental factors that shift the input distance frontier non-neutrally. The computationally simple constraint weighted bootstrapping method is employed to impose the regularity constraints on the distance function. As a by-product, total factor productivity (TFP) growth is estimated and decomposed into technical change, scale component, and efficiency change. The distance elasticities, marginal effects of the environmental factors on the distance elasticities, temporal behavior of technical efficiency, and also TFP growth and its components are investigated.

JEL Codes: D24, G21

Key Words: Input distance function; Semiparametric smooth coefficient model; Stochastic frontier model; Decomposition

1 Introduction

In estimating a production technology, both primal (e.g., a production/output distance function (ODF)) and dual (e.g., a cost/input distance function (IDF)) approaches are employed in the literature (e.g., Esho (2001), Fries and Taci (2005), Feng and Serletis (2010), Bhaumik, Das and Kumbhakar (2012), Servin, Lensink and van den Berg (2012), Sun (2015), among others). Different representations of technology have different advantages/disadvantages. For example, the production function can only handle the single output case, while the input/output distance and cost functions can accommodate multiple outputs. In fact, the ODF reduces to the production function when there is only one output. **The production function and ODF are susceptible to the endogeneity of inputs unless the constant returns to scale (CRS) restriction is imposed on them—this is because the input ratios on the right-**

*Acknowledgements: This research is funded by the National Natural Science Foundation of China (Grant ID number: 71801146). The authors would like to thank two anonymous referees for helpful comments, and remain responsible for all remaining errors.

†Corresponding author. Address: School of Economics and Finance, Curtin University, Perth WA 6845, Australia. Email: ruhul.salim@cbs.curtin.edu.au. Tel: +61 8 9266 4577.

hand-side are exogenous (Kumbhakar 2013). However, the CRS assumption is problematic since it violates the second-order conditions of profit maximization.[1]

An IDF is dual to a cost function. However, like the ODF, the estimation of the IDF does not require input price information, which is usually difficult to obtain or subject to measurement issues. The IDF gives the maximum amount by which an input vector can be radially contracted to produce the same output vector, while the ODF gives the maximum amount by which an output vector can be radially expanded using the same input vector.[2] Therefore, the IDF and ODF define input- and output-oriented technical inefficiencies, respectively. In view of the banking industry, **the IDF is preferred because it can not only handle the banks' multiple-input and multiple-output setting without input price information as the ODF, but also automatically solve the endogeneity issue of the inputs without the restrictive and problematic CRS assumption, given that the input ratios on the right-hand-side of the IDF are exogenous under cost minimization. Bank outputs (e.g., loans) are exogenous because they are constrained by consumer demand, and therefore not controlled by banks (Das and Kumbhakar 2012, Kumbhakar 2013, Gunes and Yildirim 2016).**

The estimation of technical inefficiency was pioneered by Aigner, Lovell and Schmidt (1977) and Meeusen and van den Broeck (1977) in a stochastic production frontier framework. Since then, the stochastic frontier model has been extended in many directions including the accommodation of different representations of technology as mentioned above, and also nonparametric (e.g., Fan and Li (1996) and Kumbhakar, Park, Simar and Tsionas (2007)) and semiparametric (e.g., Sun and Kumbhakar (2013), Sun, Kumbhakar and Tveterås (2015), and Yao, Zhang and Kumbhakar (2019)) specifications of the frontier function.[3] In the semiparametric specification, the traditional inputs (such as labor and capital) are in the parametric part of the regression model, while the non-traditional inputs, or environmental factors (hereafter, Z variables)—that is, firm characteristics, policy variables as well as factors that describe the environment in which production takes place—are in the nonparametric part. When all of the technology parameters (i.e., regression coefficients) are nonparametric functions of the environmental factors, the semiparametric model allows for banking heterogeneity and it introduces flexibility to the fully parametric model via non-neutral **and fully flexible** shift of the frontier of technology while maintaining some structure of the production process. This is the semiparametric smooth coefficient

[1] We would like to thank an anonymous referee for this observation.

[2] From another perspective, these representations of technologies are essentially transformation functions under different normalization restrictions (Kumbhakar and Sun 2012, Kumbhakar 2013).

[3] Although the semiparametric model is not as flexible as a fully nonparametric model, it suffers less from the "curse of dimensionality" (Li and Racine 2007).

(SPSC) model that was first proposed by Hastie and Tibshirani (1993) and Chen and Tsay (1993), and further studied by Fan and Zhang (1999), Cai, Fan and Li (2000), Li, Huang, Li and Fu (2002), Li and Racine (2010), among others. Because of its flexibility and the interpretability of the regression coefficients, the SPSC model has been widely applied in the literature (e.g. Heshmati, Kumbhakar and Sun (2014), Bhaumik, Kumbhakar and Sun (2015), among others). **It is worth noting that the advantage of the SPSC model over the random coefficient model (RCM) (Longford 1994) is that the former not only yields heterogeneous regression coefficients as the RCM does, but also generates marginal effects of the Z variables on these coefficients and on the frontier function. That is, the coefficient heterogeneity of the SPSC model is motivated by the Z variables, rather than by noises. These marginal effects quantify the extent to which a change in one of the Z variables shifts the technology frontier in a fully flexible and non-neutral manner. Under the Cobb-Douglas (CD) SPSC IDF specification, the smooth coefficients of the inputs and outputs are input and output elasticities of the IDF, and can be interpreted as cost shares of inputs and marginal costs of outputs, respectively. If Z happens to be a time trend variable, then the derivatives of the smooth coefficients of the inputs (outputs) of the IDF with respect to it—i.e., time derivatives of the input (output) elasticities—give measurements of input (scale) bias in technical change (TC) (Stevenson 1980).**[4] **If Z is banks' level of non-performing loans (NPLs), we can investigate its impact on banks' cost of production—i.e., the cost shares of inputs and marginal costs of outputs—through its impact on the input and output elasticities, respectively.**[5]

However, the price that one has to pay for the flexibility of functional form in regression setting is that the estimates may very often violate the properties dictated by economic theory. For example, the estimated marginal product or cost could possibly turn out to be negative. To overcome this potential disadvantage of the flexible regression function, regularity constraints must be imposed such that economic properties are satisfied at every data point. This is particularly important for the distance function because microeconomic theory dictates that the IDF is non-increasing in outputs, non-decreasing in inputs,

[4] **The cost shares of inputs (i.e., the input elasticities) change over time because of reallocation of resources over time, *ceteris paribus*. If banks allocate more resources in credit and portfolio analysis, then banks' output quality will improve over time. The marginal costs of outputs (i.e., the output elasticities) change over time when bank managers are risk-averse and decide to trade profit for reduced risk over time.**

[5] **As NPLs increase, a risk premium is paid to depositors who receive a higher interest rate as compensation for tolerating the extra loan risk (Hannan and Hanweck 1988, Hughes and Mester 1993). This risk premium, along with risk-averse preference of bank managers who become more reliant on deposit or short-term funding from interbank market to generate income, would affect the cost shares of inputs (i.e., the input elasticities). NPLs also trigger extra operating tasks, including additional monitoring and handling of these NPLs. These additional tasks would divert managers' attention away from normal daily monitoring of financial transactions, and therefore have implications for banks' income-generating capabilities and their marginal costs of outputs (i.e., the output elasticities).**

homogeneous of degree one in inputs, and concave in inputs. Although these economic constraints are quite important and often non-trivial to impose, none of the previous literature carried out constrained estimation for the SPSC model. In fact, the previous approaches to imposing these regularity constraints have mainly focused on fully parametric models (e.g., Diewert and Wales (1987), Terrell (1996), O'Donnell, Rambaldi and Doran (2001), O'Donnell and Coelli (2005), among others). Hall and Huang (2001) first proposed the constraint weighted bootstrapping (CWB) approach to imposing monotonicity restrictions on a univariate nonparametric model, and Du, Parmeter and Racine (2013) extended the approach to a multivariate nonparametric regression setting. Recently, Parmeter, Sun, Henderson and Kumbhakar (2014) unified these approaches to imposing constraints on a much larger class of estimators, that is, linear estimators, including the ordinary least-squares (OLS) and kernel-based nonparametric estimators. A common feature of these estimators is that they can all be expressed as weighted sums of the dependent variable. The idea of imposing constraints is to alter the dependent variable by as little as possible via a criterion function with the weights in it. **However, Parmeter et al. (2014) did not mention either the local-constant or local-linear SPSC estimator and how to impose inequality constraints on any type of the SPSC estimators at all.**

To the best of our knowledge, this is the first paper that proposes a constrained **local-linear** SPSC input distance frontier model with panel data and determinants of technical inefficiency. **We use the local-linear rather than local-constant SPSC estimator because the former simultaneously yields the smooth coefficients and their derivatives for the measurements of non-neutral effects of Z on the frontier.** Based on the observation that the **local-linear** SPSC estimator is a linear estimator, we then employ the CWB method to impose the regularity constraints **on the IDF estimated with the local-linear SPSC estimator**.[6] **Although the local-constant SPSC estimator is also a linear estimator, the constrained local-linear and local-constant SPSC estimators have different analytic forms, and it is more convenient to obtain the gradient estimates, i.e., the marginal effects of Z on the smooth coefficients, with the novel local-linear counterpart subject to the regularity constraints.** Estimation of the model is done in two steps.[7] In the first step, inefficiency is ignored and CWB is applied to estimate the constrained SPSC IDF. In the second step, the inefficiency is estimated via an auxiliary regression with certain distributional assumptions on the inefficiency and noise term. As a by-product of the estimated input distance frontier

[6]Computationally speaking, the CWB requires quadratic programming techniques which are widely available in econometric software packages.

[7]**Although the two-step estimation procedure is not efficient, it is easy for practitioners to implement, and the constrained smooth coefficients estimated in the first step do not depend on distributional assumptions. One-step estimation of the constrained local-linear SPSC model with determinants of technical inefficiency is saved for future research.**

from the first two steps, we decompose total factor productivity (TFP) growth into TC, scale component (SC), and efficiency change (EC). We then estimate and investigate the temporal behavior of technical efficiency (TE), and also TFP growth and its components. **Finally, a novel model specification test is proposed, where the null hypothesis is that the technology does not depend on Z variables. It tests the constrained local-linear SPSC model with Z versus the constrained parametric constant coefficient model without Z. The test statistic is based on comparing the residual sums of squares (RSS) of the two models, and the p-value is calculated using a nonparametric bootstrap approach (Cai, Fan and Yao 2000).**

The new methodology is applied to investigate the Indonesian banking industry with multiple inputs (i.e., total deposit, fixed assets, interest expenses and non-interest expenses), multiple outputs (i.e., total loans, other earning assets, interest income and non-interest income) and environmental factors (i.e., NPLs and time). The technology parameters (i.e., regression coefficients of the IDF) are all unknown functions of the set of environmental variables. **The application involves estimating the marginal effects of NPLs on banks' cost of production in Indonesia. NPLs are used as a measure of loan risk, which can be viewed as *ex ante* credit risk and signal future loan losses.[8] Despite minimum capital adequacy levels (i.e., capital-to-asset ratios) imposed by regulators, credit risk still seems to be one of the most threatening factors that impairs the resilience of the banking industry. Consequently the impact of credit risk measured by NPLs on bank performance seems to be an important issue to be investigated. As an integral part of the world's financial market, the assets of commercial banks in Indonesia expanded from 1038 billion Indonesian Rupiah (IDR) in 2000 to approximately 10925 billion IDR in 2015 (Bank Indonesia 2012–2017). Given the dominance of commercial banks in the Indonesian banking industry, an in-depth analysis of banking productivity and efficiency with the novel methodology and data may be interesting to researchers and policy makers.** With an unbalanced panel of 98 commercial banks in Indonesia over 16 years (i.e., from 2000 to 2015), we find that: (1) the smooth coefficient model with **NPLs and time** is preferred to a parametric model without them; (2) the constrained estimates generally have a smaller spread than the unconstrained counterpart for each smooth coefficient; (3) on average, total deposit has the largest share of cost in producing the outputs, and the marginal cost of producing interest income is higher than that of producing any of the

[8] **Financial integration that facilitates cross-border financial transactions and international capital flows allows more risk diversification, but at the same time, intensifies bank competition. As the number of loans increases, the quality of the loan portfolio is likely to decrease, which makes banks more fragile when confronted with financial crisis. While risk exposure provides banks with higher expected return, the risk management requires increased costs—banks sometimes have to trade profit for reduced risk (Hughes and Mester 1998).**

other outputs; (4) *ceteris paribus*, an increase in NPLs would increase the cost share of total deposit, increase the marginal costs of producing most outputs, and also increase the inefficiency; (5) the median TE scores generally increase over most years, indicating efficiency improvement of the Indonesian banking industry during the sample period; and finally (6) the EC and TC components almost cancel out each other for each year, and TFP growth generally follows the trajectory of the SC over the sample period. Efforts should be made in enhancing technical progress and increasing the effect of scale economy, and thus effectively boosting TFP growth and index.

The rest of this paper is organized as follows. Section 2 motivates a stochastic input distance frontier before proceeding to describe how to estimate a constrained SPSC IDF model with panel data and determinants of technical inefficiency, the decomposition of TFP growth, and a model specification test for the relevance of environmental variables. Section 3 describes the dataset. Section 4 presents and interprets the estimation results, and Section 5 concludes this paper.

2 A Semiparametric Stochastic Input Distance Frontier

Similar to the cost function, the IDF is a dual representation of technology. To derive the IDF, we follow Kumbhakar (2013) and start from the transformation function:[9]

$$A(Z) \cdot T(Y, X^*; Z) = 1, \tag{1}$$

where $A(Z) > 0$ is the productivity parameter, Z is an S-vector of environmental variables. The Z variables neutrally shift the technology frontier through the productivity parameter, $A(.)$, and non-neutrally shift the technology frontier through the transformation function, $T(.)$. $Y \in \mathbb{R}_+^K$ is a vector of the actual outputs.[10] **$X^* \in \mathbb{R}_+^J$ is a vector of minimum feasible inputs, and $X^* = X/D$, where X is a J-vector of actual inputs, $D \geq 1$ is the scalar distance by which the input vector, X, can be deflated such that it reaches X^*; therefore, $\ln D \geq 0$ is interpreted as the input-oriented technical inefficiency. The IDF is then obtained by imposing the restriction of homogeneity of degree one in X^* on the transformation function, (1), using the first optimal input, X_1^*, as the numeraire:**

$$A(Z) \cdot T(Y, \widetilde{X}^*; Z) = 1/X_1^*, \tag{2}$$

[9] See Shephard (1953, 1970) for alternative derivations of the IDF.

[10] In this paper, we focus on the input-oriented technical inefficiency; therefore, the actual outputs equal the maximum feasible outputs.

where $\widetilde{X}^*$ is a vector of input ratios, with elements $\widetilde{X}_j^* = X_j^*/X_1^* = X_j/X_1 = \widetilde{X}_j$, $\forall j = 2, \ldots, J$. Using the fact that $X_1^* = X_1/D$, the IDF (2) can be rewritten as:

$$D/X_1 = A(Z) \cdot T(Y, \widetilde{X}; Z) = A(Z) \prod_{k=1}^{K} Y_k^{\gamma_k(Z)} \prod_{j=2}^{J} \widetilde{X}_j^{\beta_j(Z)}, \tag{3}$$

where $\widetilde{X}$ is a vector of input ratios, with elements $\widetilde{X}_j$, $\forall j = 2, \ldots, J$. The returns to scale, $RTS = -1/\sum_{k=1}^{K} \gamma_k(Z)$.

Taking the natural logarithm for both sides and adding a noise term, v, gives:

$$\begin{aligned} -\ln X_1 &= \ln A(Z) + \ln T(Y, \widetilde{X}; Z) + v - \ln D \\ &= \alpha(Z) + \sum_{k=1}^{K} \gamma_k(Z) \ln Y_k + \sum_{j=2}^{J} \beta_j(Z) \ln \widetilde{X}_j + v - u, \end{aligned} \tag{4}$$

where $\alpha(Z) = \ln A(Z)$. The inefficiency is defined as $u = \ln D \geq 0$, since $D \geq 1$ is the value of the IDF. Note that the $v - u$ term in (4) is analogous to the composite error in a production frontier. Based on this observation, we can then estimate the input distance frontier, (4), using exactly the same procedure as we estimate the production frontier.

The properties of the IDF implies that:

$$\frac{\partial \ln D}{\partial \ln Y_k} \leq 0, \forall k = 1, \ldots, K; \text{ and} \tag{5}$$

$$\frac{\partial \ln D}{\partial \ln X_j} \geq 0, \forall j = 1, \ldots, J. \tag{6}$$

Using the fact that $D = X_1 \cdot A(Z) \cdot T(Y, \widetilde{X}; Z)$, or equivalently, $\ln D = \ln X_1 + \ln A(Z) + \ln T(Y, \widetilde{X}; Z)$, we can rewrite these constraints for ease of estimation as:

$$\frac{\partial \ln D}{\partial \ln Y_k} = \frac{\partial \ln T(Y, \widetilde{X}; Z)}{\partial \ln Y_k} = \gamma_k(Z) \leq 0, \forall k = 1, \ldots, K, \tag{7}$$

$$\frac{\partial \ln D}{\partial \ln X_1} = 1 - \sum_{j=2}^{J} \frac{\partial \ln T(Y, \widetilde{X}; Z)}{\partial \ln \widetilde{X}_j} = 1 - \sum_{j=2}^{J} \beta_j(Z) \equiv \beta_1(Z) \geq 0, \text{ and} \tag{8}$$

$$\frac{\partial \ln D}{\partial \ln X_j} = \frac{\partial \ln T(Y, \widetilde{X}; Z)}{\partial \ln \widetilde{X}_j} = \beta_j(Z) \geq 0, \forall j = 2, \ldots, J. \tag{9}$$

Note that the functional form of $T(.)$ in (3) has the CD structure with flexible parameters—distance elasticities, γ and β—as unknown functions of Z. These elasticities measure the sensitivity of a producer's

radial distance to the production possibility frontier to a change in output or input quantities. Thus, the functional form of (3) is more flexible than the traditional CD specification since (3) generates parameter heterogeneity of γ and β; however, (3) preserves the interpretability of its parameters—that is, we can interpret the γ and β in (3) in the same way as we interpret them in the traditional CD specification without the Z variables. **In addition, given that the distance elasticities, γ and β in** (3) **do not change with output and input quantities, the CD structure in** (3) **guarantees that given Z, the input requirement set of the IDF is convex, and therefore the IDF is concave in inputs.**

2.1 Constrained Semiparametric Smooth Coefficient Estimation

To estimate the regression coefficients and technical inefficiency in (4), we follow Sun and Kumbhakar (2013) and Yao et al. (2019) and use a two-step estimation procedure. In the first step, we transform the stochastic frontier model (4) and estimate the regression coefficients subject to the constraints specified in (7)–(9).[11] The technical inefficiency is then estimated in the second step in which a panel stochastic frontier model is estimated. **More specifically, the composite error term is $\varepsilon_{it}^{**} \equiv v_{it} - u_{it}$, $\forall i = 1, \ldots, N$ and $t = 1, \ldots, T$. Since u_{it} is non-negative, it is obvious that $E(\varepsilon_{it}^{**} | \ln Y_{it}, \ln \widetilde{X}_{it}, Z_{it})$ is non-zero.[12] Following Wang and Ho (2010), we assume that $E(\varepsilon_{it}^{**} | \ln Y_{it}, \ln \widetilde{X}_{it}, Z_{it}) = E(\varepsilon_{it}^{**} | Z_{it})$; that is, conditional on Z_{it}, ε_{it}^{**} is mean independent from $\ln Y_{it}$ and $\ln \widetilde{X}_{it}$. Given that $E(v_{it} | \ln Y_{it}, \ln \widetilde{X}_{it}, Z_{it}) = E(v_{it} | Z_{it}) = 0$, we would have $E(\varepsilon_{it}^{**} | \ln Y_{it}, \ln \widetilde{X}_{it}, Z_{it}) = -E(u_{it} | \ln Y_{it}, \ln \widetilde{X}_{it}, Z_{it}) = -E(u_{it} | Z_{it})$. To fix the issue of non-zero conditional mean of ε_{it}^{**}, let $\varepsilon_{it} \equiv \varepsilon_{it}^{**} + E(u_{it} | Z_{it})$. It then becomes obvious that $E(\varepsilon_{it} | \ln Y_{it}, \ln \widetilde{X}_{it}, Z_{it}) = E(\varepsilon_{it}^{**} | \ln Y_{it}, \ln \widetilde{X}_{it}, Z_{it}) + E(u_{it} | \ln Y_{it}, \ln \widetilde{X}_{it}, Z_{it}) = 0$. Equivalently,** (4) **can be rewritten as:**

$$-\ln X_{1it} = \widetilde{\alpha}(Z_{it}) + \sum_{k=1}^{K} \gamma_k(Z_{it}) \ln Y_{kit} + \sum_{j=2}^{J} \beta_j(Z_{it}) \ln \widetilde{X}_{jit} + \varepsilon_{it}, \tag{10}$$

where $\widetilde{\alpha}(Z_{it}) = \alpha(Z_{it}) - E(u_{it} | Z_{it})$.[13] (10) **can be viewed as a standard SPSC model (Li et al. 2002). Define a scalar $\mathcal{Y}_{it} = -\ln X_{1it}$, two $(K+J)$-vectors $\rho(Z_{it}) = [\widetilde{\alpha}(\cdot), \gamma_1(\cdot), \ldots, \gamma_K(\cdot), \beta_2(\cdot), \ldots, \beta_J(\cdot)]$**

[11] Neither Sun and Kumbhakar (2013) nor Yao et al. (2019) mentioned how to estimate the SPSC model subject to inequality constraints in their papers.

[12] Unlike traditional fully parametric stochastic frontier models in which Z_{it} only appears in the inefficiency function rather than the frontier function, the SPSC stochastic frontier models include Z_{it} in the inefficiency and frontier functions at the same time. Therefore, the estimating equation, $E(\varepsilon_{it}^{**} | \cdot)$ must include Z_{it}, in addition to $\ln Y_{it}$ and $\ln \widetilde{X}_{it}$.

[13] That is, the $E(u_{it} | Z_{it})$ is absorbed by the intercept, $\widetilde{\alpha}(Z_{it})$.

and $W_{it} = [1, \ln Y_{1it}, \ldots, \ln Y_{Kit}, \ln \widetilde{X}_{2it}, \ldots, \ln \widetilde{X}_{Jit}]$**, respectively, and** (10) **can be written as:**

$$\mathcal{Y}_{it} = W_{it}' \rho(Z_{it}) + \varepsilon_{it}. \tag{11}$$

We use the local-linear SPSC estimator to estimate (11)**, given that this estimator simultaneously yields the smooth coefficients and the derivatives of the smooth coefficients with respect to** Z_{it}**. Let** $\widetilde{W}_{\iota\tau} = \begin{bmatrix} W_{\iota\tau} \\ W_{\iota\tau} \bigotimes (Z_{\iota\tau} - Z_{it}) \end{bmatrix}$ **be a** $[(K+J) \times (1+S)]$**-vector, and** $\tilde{\rho}(Z_{it}) = [\rho(Z_{it})' \ \nabla\rho(Z_{it})']'$ **be a** $[(K+J) \times (1+S)]$**-vector of parameters, where** $\bigotimes$ **denotes the Kronecker product, and** ∇ **denotes the gradient vector of** $\rho(Z_{it})$ **with respect to** Z_{it}**. The local-linear SPSC estimator of** $\tilde{\rho}(Z_{it})$ **is then the solution of** $\hat{\tilde{\rho}}(Z_{it})$ **to the locally-weighted sample moment condition:** $\sum_{\iota=1}^{N} \sum_{\tau=1}^{T} \widetilde{W}_{\iota\tau} (\mathcal{Y}_{\iota\tau} - \widetilde{W}_{\iota\tau}' \hat{\tilde{\rho}}(Z_{it})) \mathcal{K}_h(Z_{\iota\tau}, Z_{it}) = 0$**, where** $\mathcal{K}_h(Z_{\iota\tau}, Z_{it}) = \prod_{s=1}^{S} \mathcal{K}_s(\cdot)$**,** $\mathcal{K}_s(\cdot) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{Z_{s,\iota\tau} - Z_{s,it}}{h_s}\right)^2\right)$ **denotes the Gaussian kernel function for each environmental variable, and** h_s **is the bandwidth for the** s**th variable in** Z_{it}**. It is easy to show that:**[14]

$$\hat{\tilde{\rho}}(Z_{it}) = [\hat{\rho}(Z_{it})' \ \nabla\hat{\rho}(Z_{it})']' = \left[\sum_{\iota=1}^{N} \sum_{\tau=1}^{T} \widetilde{W}_{\iota\tau} \widetilde{W}_{\iota\tau}' \mathcal{K}_h(Z_{\iota\tau}, Z_{it})\right]^{-1} \sum_{\iota=1}^{N} \sum_{\tau=1}^{T} \widetilde{W}_{\iota\tau} \mathcal{Y}_{\iota\tau} \mathcal{K}_h(Z_{\iota\tau}, Z_{it}). \tag{12}$$

Because ρ is an unknown function of Z_{it}, $\hat{\rho}(Z_{it})$ is observation-specific; viz., there is a parameter vector ρ given each observation of Z_{it}. One would expect that the theoretical properties of the IDF (i.e., (7)–(9)) are satisfied for each observation. However, in empirical work, when the functional form is flexible or unknown, one may find violations of the theoretical properties for a number of observations in a dataset. This may be viewed as the price that has to be paid for the relaxation of the functional form assumption. To overcome this problem, we propose a constrained SPSC model, where we are able to guarantee that all the estimated coefficients (or a linear combination of them) in (12) have economic meaning.[15] To do this, we rewrite (12) as:

$$\hat{\tilde{\rho}}(Z_{it}) = \sum_{\iota=1}^{N} \sum_{\tau=1}^{T} A_{\iota\tau}(\widetilde{W}_{\iota\tau}, Z_{\iota\tau}, Z_{it}) \mathcal{Y}_{\iota\tau}, \tag{13}$$

[14] The bandwidth selection method for estimating $\rho(\cdot)$ is the least-squares cross-validation (LSCV). Appendix A explains in detail how we select the bandwidths to estimate $\tilde{\rho}(\cdot)$.

[15] It is not necessary to test for the validity of the economic constraints given in (7)–(9) that are dictated by economic theory. For example, we do not want to test if the non-negative marginal cost constraint is valid when estimating a cost function. This is because the non-negative marginal cost is one of the properties of the cost function dictated by microeconomic theory. Data points that violate such a property would not be of any interest to policy makers. Test for the validity of constraints is warranted when some non-economic constraints are imposed—see Du et al. (2013) for testing for the validity of constraints under the CWB framework via bootstrapping.

where $A_{\iota\tau}(\cdot) = \left[\sum_{\iota=1}^{N}\sum_{\tau=1}^{T}\widetilde{W}_{\iota\tau}\widetilde{W}'_{\iota\tau}\mathcal{K}_h(Z_{\iota\tau}, Z_{it})\right]^{-1}\widetilde{W}_{\iota\tau}\mathcal{K}_h(Z_{\iota\tau}, Z_{it})$. The idea of imposing the observation-specific constraints on an SPSC estimator is simply reweighting each observation of the dependent variable, $\mathcal{Y}_{\iota\tau}$ (Hall and Huang 2001, Du et al. 2013). To do this, we rewrite (13) as:

$$\hat{\tilde{\rho}}(Z_{it}) = NT \cdot \sum_{\iota=1}^{N}\sum_{\tau=1}^{T} A_{\iota\tau}(\widetilde{W}_{\iota\tau}, Z_{\iota\tau}, Z_{it}) \cdot p_u \cdot \mathcal{Y}_{\iota\tau}, \tag{14}$$

where $p_u = (NT)^{-1}$ denotes the uniform weights. **The unconstrained local-linear SPSC estimator is given in** (14)**. To impose the constraints, we can write the constrained estimator as:**

$$\hat{\tilde{\rho}}^*(Z_{it}) = NT \cdot \sum_{\iota=1}^{N}\sum_{\tau=1}^{T} A_{\iota\tau}(\widetilde{W}_{\iota\tau}, Z_{\iota\tau}, Z_{it}) \cdot p_{\iota\tau} \cdot \mathcal{Y}_{\iota\tau}, \tag{15}$$

where $\hat{\tilde{\rho}}^*(Z_{it}) = [\hat{\rho}^*(Z_{it})' \ \nabla\hat{\rho}^*(Z_{it})']'$ **denotes the constrained local-linear SPSC estimator**, $p_{\iota\tau}$ denotes the observation-specific weights, and $\sum_{\iota=1}^{N}\sum_{\tau=1}^{T} p_{\iota\tau} = 1$. To select the optimal $p_{\iota\tau}$, we follow Du et al.'s (2013) approach and minimize the L_2-norm criterion function:

$$\begin{aligned} &\sum_{\iota=1}^{N}\sum_{\tau=1}^{T}(p_{\iota\tau} - p_u)^2 \\ &\text{subject to } \hat{\gamma}_k(Z_{it}) \leq 0, \forall k = 1, \ldots, K; \ 1 - \sum_{j=2}^{J}\hat{\beta}_j(Z_{it}) \geq 0; \text{ and} \\ &\hat{\beta}_j(Z_{it}) \geq 0, \forall j = 2, \ldots, J. \end{aligned} \tag{16}$$

This is a quadratic programming procedure, **and the constraints must be internally consistent with each other for the solution to the optimal weights to be feasible**. We use the *quadprog* package in R to solve for the optimal $p_{\iota\tau}$. We observe some economic violations in our empirical application, and then apply the proposed constrained estimation approach to imposing the constraints on each smooth coefficient and a linear combination of them.[16] As a by-product, we can also obtain the marginal effects of each variable in Z_{it} on $\mathcal{Y}_{it}$ through its marginal effects on $\hat{\rho}^*(Z_{it})$, i.e., $\nabla\hat{\rho}^*(Z_{it})$.

2.2 A Non-linear Stochastic Frontier Model

Recall from Section 2.1 that $\varepsilon_{it} \equiv E(u_{it}|Z_{it}) + v_{it} - u_{it}$. We then follow Yao et al. (2019) and assume that (1) $v_{it} \sim iidN(0, \sigma_v^2)$ is a random shock, and (2) $u_{it} = u_i g(Z_{it}; \eta)$, where $u_i \sim iidN^+(0, \sigma_u^2)$ is independent from Z_{it} and is called base inefficiency level, $g(Z_{it}; \eta) = \exp(\eta' Z_{it})$ is the scaling function (Wang and Schmidt 2002), and η is an S-vector. Therefore, it follows that $u_{it} \sim iidN^+(0, \sigma_u^2 g^2(Z_{it}; \eta))$,

[16] The R codes for imposing these constraints are available from the authors upon request.

and $E(u_{it}|Z_{it}) = \sqrt{2/\pi}\sigma_u g(Z_{it};\eta)$.

To estimate the parameters, i.e., η, σ_u^2, and σ_v^2, we would need to apply a standard non-linear stochastic frontier approach; that is,

$$\begin{aligned}\hat{\varepsilon}_{it}^{*} &= E(u_{it}|Z_{it}) + v_{it} - u_{it} \\ &= \sqrt{2/\pi}\sigma_u g(Z_{it};\eta) + v_{it} - u_{it} \\ &= \sqrt{2/\pi}\sigma_u \exp(\eta' Z_{it}) + v_{it} - u_{it},\end{aligned} \tag{17}$$

where $\hat{\varepsilon}_{it}^{*} = \mathcal{Y}_{it} - W_{it}'\hat{\rho}^{*}(Z_{it})$ are the residuals estimated from (11) using the constrained estimation method, and use the maximum likelihood estimation method. Recall that $\varepsilon_{it}^{**} \equiv v_{it} - u_{it}$, the log-likelihood function for (17) is (Yao et al. 2019):

$$\ln L \propto -\frac{N(T-1)}{2}\ln\sigma_v^2 - \frac{1}{2}\sum_{i=1}^{N}\ln\sigma_i^2 + \sum_{i=1}^{N}\ln\Phi\left(\frac{\mu_{*i}}{\sigma_{*i}}\right) + \frac{1}{2}\sum_{i=1}^{N}\left(\frac{\mu_{*i}}{\sigma_{*i}}\right)^2 - \frac{1}{2\sigma_v^2}\sum_{i=1}^{N}\sum_{t=1}^{T}\varepsilon_{it}^{**2}, \tag{18}$$

where $\sigma_i^2 = \sigma_v^2 + \sigma_u^2\sum_{t=1}^{T} g^2(Z_{it};\eta)$, $\mu_{*i} = -\sigma_u^2\sum_{t=1}^{T}\varepsilon_{it}^{**}g(Z_{it};\eta)/\sigma_i^2$, and $\sigma_{*i}^2 = \sigma_u^2\sigma_v^2/\sigma_i^2$. Maximizing (18) with respect to η, σ_u^2 and σ_v^2 gives the maximum likelihood estimates (MLE) of these parameters, which are then used to estimate $E(u_{it}|Z_{it})$. The original intercept can then be recovered using $\alpha(Z_{it}) = \widetilde{\alpha}(Z_{it}) + E(u_{it}|Z_{it})$.

The point estimator of u_i, $\forall i = 1,\ldots,N$, is:

$$E(u_i|\varepsilon_i^{**}, Z_i) = \mu_{*i} + \sigma_{*i}\frac{\phi(\mu_{*i}/\sigma_{*i})}{\Phi(\mu_{*i}/\sigma_{*i})}. \tag{19}$$

Given that $u_{it} = u_i g(Z_{it};\eta)$, technical inefficiency of u_{it} is then computed from $\hat{E}(u_i|\varepsilon_i^{**}, Z_i)g(Z_{it};\hat{\eta})$, and the TE score is $\exp(-\hat{E}(u_i|\varepsilon_i^{**}, Z_i)g(Z_{it};\hat{\eta}))$, $\forall i = 1,\ldots,N$ and $t = 1,\ldots,T$.

Finally, inefficiency changes ($\partial u_{it}/\partial t$, $\exists t \in Z$) are calculated as:

$$\frac{\partial u_{it}}{\partial t} = \frac{\partial \hat{E}(u_i|\varepsilon_i^{**}, Z_i)g(Z_{it};\hat{\eta})}{\partial t} = \hat{E}(u_i|\varepsilon_i^{**}, Z_i)\exp(\hat{\eta}' Z_{it})\hat{\eta}_t, \tag{20}$$

where $\hat{\eta}_t \in \hat{\eta}$ is the coefficient associated with $t \in Z_{it}$ in $g(Z_{it};\hat{\eta})$.

2.3 TFP Growth and Its Components

Using the estimated smooth coefficients **and their derivatives**, we can further estimate TFP growth and its components. To see this, we use the Divisia measure of TFP growth; that is, $T\dot{F}P = \sum_{k=1}^{K} R_k\dot{Y}_k - \sum_{j=1}^{J} S_j\dot{X}_j$, where R_k denotes the revenue share of each output ($k = 1, ..., K$) and S_j represents the cost

share of each input ($j = 1, ..., J$). **Under perfect competition,**

$$S_j = \beta_j(\cdot) = \frac{\partial \ln D}{\partial \ln X_j}, \ \forall j = 1, ..., J, \tag{21}$$

is the shadow cost share for the jth input, and

$$R_k = \frac{\gamma_k(\cdot)}{\sum_{k=1}^{K} \gamma_k(\cdot)} = \frac{\partial \ln D / \partial \ln Y_k}{\sum_{k=1}^{K} \partial \ln D / \partial \ln Y_k}, \ \forall k = 1, ..., K, \tag{22}$$

is the shadow revenue share for the kth output.

For the ease of decomposition, we take the time derivatives of both sides of (4) **(ignoring the noise term), and then add** $T\dot{F}P$ **to both sides to obtain** $T\dot{F}P =$ Technical Change $+$ Scale Component $+$ Allocative Component $+$ Efficiency Change,[17] where

$$\text{Technical Change (TC)} = \frac{\partial \alpha(Z)}{\partial t} + \sum_{k=1}^{K} \frac{\partial \gamma_k(Z)}{\partial t} \ln Y_k + \sum_{j=2}^{J} \frac{\partial \beta_j(Z)}{\partial t} \ln \widetilde{X}_j, \ \exists t \in Z \tag{23}$$

where t is the time trend variable.

$$\begin{aligned} \text{Scale Component (SC)} &= \sum_{k=1}^{K} (R_k + \gamma_k(Z)) \dot{Y}_k \\ &= (1 - RTS) \sum_{k=1}^{K} \gamma_k(Z) \dot{Y}_k + \sum_{k=1}^{K} \left(R_k - \frac{\gamma_k(Z)}{\sum_{k=1}^{K} \gamma_k(Z)} \right) \dot{Y}_k. \end{aligned} \tag{24}$$

The SC is further decomposed into two sub-components. The first sub-component depends on RTS and equals zero under CRS (i.e., when $RTS = 1$), and the second sub-component equals zero by (22). The Allocative Component (AC) $= \sum_{j=2}^{J} (\beta_j(Z) - S_j) \dot{\widetilde{X}}_j = 0$ by (21). Finally, the Efficiency Change (EC) component is given by $-\partial u / \partial t$.

2.4 Testing for the relevance of environmental variables

To ascertain that the technology depends on Z **variables, we can test if** (4) **can be estimated as a standard parametric stochastic frontier model without the** Z **variables:**

$$-\ln X_1 = \alpha + \sum_{k=1}^{K} \gamma_k \ln Y_k + \sum_{j=2}^{J} \beta_j \ln \widetilde{X}_j + \nu - \mu, \tag{25}$$

[17] We use the fact that $\sum_{j=1}^{J} S_j = 1$ and $\dot{\widetilde{X}}_1 = 0$. See Appendix B for derivation details for the decomposition. **Kumbhakar and Wang (2005) showed a similar decomposition based on a parametric production function.**

where ν is the iid normal noise term with $E(\nu|\ln Y, \ln \widetilde{X}) = 0$, and μ is the iid half-normal technical inefficiency term with $E(\mu|\ln Y, \ln \widetilde{X}) = E(\mu)$. The technology in (25) is independent of Z, because neither the coefficients nor technical inefficiency depends on Z. Equivalently, (25) can be rewritten as:

$$-\ln X_1 = \tilde{\alpha} + \sum_{k=1}^{K} \gamma_k \ln Y_k + \sum_{j=2}^{J} \beta_j \ln \widetilde{X}_j + \epsilon, \tag{26}$$

where $\tilde{\alpha} = \alpha - E(\mu)$, and $\epsilon = \nu - \mu + E(\mu)$. It is obvious that $E(\epsilon|\ln Y, \ln \widetilde{X}) = E(\nu|\ln Y, \ln \widetilde{X}) - E(\mu|\ln Y, \ln \widetilde{X}) + E(E(\mu)|\ln Y, \ln \widetilde{X}) = 0$. Let $\mathcal{Y} = -\ln X_1$, $W = [1, \ln Y_1, \ldots, \ln Y_K, \ln \widetilde{X}_2, \ldots, \ln \widetilde{X}_J]$, $\rho = [\tilde{\alpha}, \gamma_1, \ldots, \gamma_K, \beta_2, \ldots, \beta_J]$, and add subscripts i and t to the variables in (26), and (26) can be rewritten as:

$$\mathcal{Y}_{it} = W_{it}'\rho + \epsilon_{it}. \tag{27}$$

The unconstrained OLS estimator of $\hat{\rho}$ is:

$$\hat{\rho} = \left(\sum_{i=1}^{N} \sum_{t=1}^{T} W_{it} W_{it}' \right)^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} W_{it} \mathcal{Y}_{it}. \tag{28}$$

To estimate $\hat{\rho}$ subject to the regularity constraints on the input and output elasticities,[18] rewrite the unconstrained OLS estimator of $\hat{\rho}$ as:

$$\hat{\rho} = NT \cdot \sum_{i=1}^{N} \sum_{t=1}^{T} A_{it} \cdot p_u \cdot \mathcal{Y}_{it}, \tag{29}$$

where $A_{it} = (\sum_{i=1}^{N} \sum_{t=1}^{T} W_{it} W_{it}')^{-1} W_{it}$, and $p_u = (NT)^{-1}$ denotes the uniform weights. The constrained OLS estimator then becomes:

$$\hat{\rho}^* = NT \cdot \sum_{i=1}^{N} \sum_{t=1}^{T} A_{it} \cdot p_{it} \cdot \mathcal{Y}_{it}, \tag{30}$$

where p_{it} denotes the observation-specific weights, and $\sum_{i=1}^{N} \sum_{t=1}^{T} p_{it} = 1$. It is worth noting that the SPSC estimator is the OLS estimator conditional on Z. Indeed, the $\hat{\tilde{\rho}}(Z_{it})$ in (12) becomes $\hat{\rho}$ in (28) when $\widetilde{W}_{\iota\tau}$ is replaced by W_{it}, and $\mathcal{K}_h(Z_{\iota\tau}, Z_{it})$ is replaced by 1. The same

[18] **See Parmeter et al. (2014) for more details about imposing constraints on fully parametric models via the CWB approach.**

criterion function is used to select the optimal weights, p_{it}, i.e., minimize

$$\sum_{i=1}^{N}\sum_{t=1}^{T}(p_{it}-p_u)^2$$
$$\text{subject to } \hat{\gamma}_k \leq 0, \forall k=1,\ldots,K;\ 1-\sum_{j=2}^{J}\hat{\beta}_j \geq 0;\ \text{and } \hat{\beta}_j \geq 0, \forall j=2,\ldots,J. \tag{31}$$

The technology in (25) **is independent of** Z **if** ρ **is a constant, i.e., does not vary with** Z**. Therefore, the null hypothesis to be tested is:** $H_0 : \rho(Z_{it}) = \rho$**. Following Cai, Fan and Yao (2000), we test the null hypothesis by comparing the RSS of the constant coefficient and SPSC models. The RSS under** H_0**, i.e., the constant coefficient model, is** $RSS_0 = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\hat{\epsilon}_{it}^{*2}$**, where** $\hat{\epsilon}_{it}^{*} = \mathcal{Y}_{it} - W_{it}'\hat{\rho}^{*}$ **are the residuals of the constrained OLS model. Similarly, the RSS under the SPSC model is** $RSS_1 = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\hat{\varepsilon}_{it}^{*2}$**, where** $\hat{\varepsilon}_{it}^{*} = \mathcal{Y}_{it} - W_{it}'\hat{\rho}^{*}(Z_{it})$ **are the residuals of the constrained SPSC model. The test statistic is then constructed as:**

$$T_n = \frac{RSS_0 - RSS_1}{RSS_1} = \frac{RSS_0}{RSS_1} - 1. \tag{32}$$

To obtain the p**-value of the test and determine whether to reject the null hypothesis or not, we follow Cai, Fan and Yao's (2000) nonparametric bootstrap approach with the following steps:**

Step 1: generate bootstrap residuals, ε_{it}^{b}**, using the centralized residuals from the SPSC model, i.e.,** $\hat{\varepsilon}_{it}^{*} - \bar{\hat{\varepsilon}}_{it}^{*}$**, where** $\bar{\hat{\varepsilon}}_{it}^{*}$ **is the sample mean of** $\hat{\varepsilon}_{it}^{*}$**;**

Step 2: generate $\mathcal{Y}_{it}^{b} = W_{it}'\hat{\rho}^{*} + \varepsilon_{it}^{b}$**;**

Step 3: use $\{\mathcal{Y}_{it}^{b}, W_{it}, Z_{it}\}_{i=1,\ldots,N;\, t=1,\ldots,T}$ **to calculate the bootstrap test statistic,** T_n^{b}**;**

Step 4: repeat the previous steps a large number of times, say, $B = 399$**, and the** p**-value is calculated as** $\frac{1}{B}\sum_{b=1}^{B}\mathbf{I}(T_n^{b} > T_n)$**, where** $\mathbf{I}(\cdot)$ **is the indicator function with a value of 1 if its argument is true. The null hypothesis can be rejected if the** p**-value is less than the level of significance, say, 0.05.**

3 Data

The data used in this study are sourced from banks' financial statements, which are annually published by the Central Bank of Indonesia (i.e., Bank Indonesia) over the 16-year period from 2000–2015. This data period is chosen on the basis of the availability of consistent data. We included banks in continuous

operation in this sample period. However, some banks with missing data are excluded[19] from our sample, and thus we have a sample of 98 banks comprising of four state banks, 51 private banks, 25 regional development banks, 10 joint venture banks and eight foreign banks. Our sample covers over 80% of total commercial bank assets of Indonesia during the sample period. All of the input and output variables are measured in IDR and are deflated by the GDP deflator. **In the literature, the two frequently used approaches to specifying bank's input and output variables are the intermediation and production approaches, respectively. Neoclassical microeconomic theory is the underlying basis for both approaches; however, they differ only in the specification of banking activities (Das and Kumbhakar 2012). The intermediation approach, introduced by Sealey and Lindley (1977), considers banks as mediators of funds from depositors to borrowers, with deposits used to produce loans and other assets; while the production approach treats banks as the provider of services to customers or production center, where banks utilize physical inputs (e.g., labor and capital) to produce deposits and other outputs (Denizer 2000). Both approaches are equally popular in the literature; however, Berger and Humphrey (1997) advocated that the intermediation approach be used for measuring efficiency for the entire financial institutions, and production approach for the bank branch level. In this study, we follow Defung, Salim and Bloch (2016) and adopt the intermediation approach to constructing input and output variables for the Indonesian Banks. Drawing on the previous theoretical and empirical studies (Berger and DeYoung 1997, Fofack 2005, Park and Weber 2006, Assaf, Matousek and Tsionas 2013) and the availability of data, we identify NPLs and time as the environmental variables that not only shift the distance frontier in non-neutral manners, but also determine the level of inefficiency**, since the technology parameters or efficiency scores alone have limited utility for bank management and policy purposes unless the applied research investigates the factors affecting these parameters and efficiency scores. Analysis of the impacts of these environmental variables on banks' productivity and efficiency has become even more important in the wake of banks' failures in many countries during and after the global financial crisis (Podpiera and Weill 2008). Given the market and production conditions in Indonesia, it can be argued that time can promote bank efficiency, whereas NPLs can worsen efficiency, *ceteris paribus*. We will discuss how these environmental variables shape up the frontier in more details in the following section. Summary statistics of the inputs, outputs and environmental factors are presented in Table 1.

[19]Banks liquidated or closed down, or banks established during the sample period, are also excluded from our dataset.

4 Estimation Results

Our estimation results are presented in Figures 1–6 and Tables 2–4. Figure 1 plots the kernel density functions of the smooth coefficient estimates for the unconstrained and constrained models with dashed and solid lines, respectively. Silvermans reflection method is employed for the constrained distributions because the underlying histograms may have masses near zero. It is clear that the constrained estimates deviate from their unconstrained counterpart, and that the estimates that were of the wrong sign are not simply set to zero. Unsurprisingly, the constrained estimates generally have smaller spread than the unconstrained counterpart for each smooth coefficient. In fact, with respect to the smooth coefficient estimates with lower percentages of violation, the constrained and unconstrained models do not differ as much as when such percentages are higher. For example, we can see a 'spike' at zero for β_2, the smooth coefficient with the highest percentage of violations, and actually many estimates with violations move just to the other side of the zero vertical line with a small shift to the left in the other parts of its distribution.

Table 2 reports the mean and quartile values (Q1–Q3) of the estimated smooth coefficients of interest for both the unconstrained and constrained models, **along with bootstrap standard errors[20] in the parentheses and asymptotic standard errors[21] in the brackets. It can be seen that the bootstrap standard errors are similar to their asymptotic counterparts in Table 2.** Percentages of violations are reported under the unconstrained model. It can be seen that a non-trivial number of observations violate the theoretical properties of the distance function for β_2. These violations have motivated us to estimate the constrained counterpart, and the results from the constrained model satisfying economic theory are more credible and should be used to estimate the TE in the non-linear stochastic frontier model. To verify the importance of estimating the constrained model, the Kolmogorov-Smirnov tests for the null of equality of distributions across the unconstrained and constrained estimates are conducted and the p-values are reported under the constrained model. It can be seen that the null of equal distributions with a two-sided alternative can be rejected at the 1% level for all the smooth coefficients.[22]

[20] **The bootstrap standard error of the mean (quartile) of each quantity, say, γ_1, is calculated as follows (Cameron and Trivedi 2005, Chapter 11). First, generate a bootstrap sample γ_1^b by randomly selecting from γ_1 with replacement. Second, compute the mean (quartile) of γ_1^b. Call it the bootstrap mean (quartile) estimate. Repeat these two steps 1000 times, and the standard error of the mean (quartile) is viewed as the standard deviation calculated using the bootstrap mean (quartile) estimates.**

[21] **The asymptotic standard errors of the mean are calculated by dividing the sample standard deviation of each quantity, say, γ_1, by the square root of the sample size, and those of the quartiles are calculated according to Koenker and Bassett (1978). The *quantreg* package of R is useful for this purpose.**

[22] Even if there is no violation for a particular quantity (e.g., γ_3 and β_4), it is still possible that the unconstrained and constrained densities significantly differ from each other. This is because a large number of violations of any other quantities (e.g., β_2) would require a significant amount of perturbation of the uniform weights, and hence the dependent variable. In

Given that it is non-trivial to impose the constraints dictated by the properties of the IDF, we focus on the constrained model when interpreting the estimation results. In particular, β_j **is the distance elasticity with respect to the** j**th input,** $\forall j$**, as can be seen from** (8) **and** (9)**.**[23] **It can be interpreted as the (shadow) cost share of the** j**th input. The homogeneity property of the IDF indicates that the sum of all the** β_j**'s (i.e.,** $\sum_{j=1}^{J} \beta_j$**) must equal unity. The constrained section of Table 2 shows that, on average, total deposit** (X_1) **has the largest share of cost in producing the outputs—the mean** β_1 **is 0.5472; that is, cost of total deposit accounts for 54.72% of the total (shadow) cost in producing the four outputs. Alternatively, a 1% increase in total deposit would cause the total (shadow) cost to increase by 0.5472%,** ***ceteris paribus.*** **This result is expected as deposit-taking is one of the leading activities of Indonesian banks. On the contrary, fixed assets** (X_2) **has the smallest share of cost in producing the outputs—the mean** β_2 **is 0.0143; that is, cost of fixed assets only accounts for 1.43% of the total cost, and a 1% increase in fixed assets would cause the total cost to increase by 0.0143%,** ***ceteris paribus.*** **This result supports the idea that large banks in Indonesia are relatively cost efficient, and conforms to that in Hadad, Hall, Santoso, Satria, Kenjegalieva and Simper (2008).**

γ_k **can be interpreted as the distance elasticity with respect to the** k**th output,** $\forall k$**. Alternatively,** $-\gamma_k$ **shows the percent increase in the inputs**[24] **due to a 1% increase in the production of** Y_k**. Therefore, in economics terms,** $-\gamma_k$ **measures the opportunity cost of producing one more percent of the** k**th output; that is, it is an equivalence of marginal cost in a cost function. It can be seen that, on average, it is most costly to produce an additional percent of interest income** (Y_3)**—the mean** $-\gamma_3$ **is 0.5448; that is, a 1% increase in interest income would cause the total cost to increase by 0.5448%,** ***ceteris paribus.*** **On the contrary, it is least costly to produce an additional percent of non-interest income** (Y_4)**—the mean** $-\gamma_4$ **is 0.0703; that is, a 1% increase in non-interest income would cause the total cost to increase by only 0.0703%,** ***ceteris paribus.*** **The** RTS **is calculated as** $-1/\sum_{k=1}^{K} \gamma_k(Z)$**, and the mean RTS estimate is 0.9987 with a standard deviation of 0.0283. The RTS estimates range from 0.9188 to 1.0989—there is a slight degree of heterogeneity in** RTS **in the banks in Indonesia. In general, this heterogeneity is due to different bank sizes and ownership**

summary, it is not that a smaller percentage of violations must produce a larger p-value, it is how the weights have to change to ensure that all the constraints are imposed simultaneously.

[23] Recall that $\beta_1 \equiv \partial \ln D / \partial \ln X_1 = 1 - \sum_{j=2}^{J} \beta_j$.

[24] **This is because a 1% increase in** Y_k **would cause** X_1 **to increase by** $-\gamma_k$**%, and in order to hold everything else constant, including the input ratios in log (i.e.,** $\ln X_j - \ln X_1$**),** X_j **must also increase by the same percentage, i.e.,** $-\gamma_k$**%,** $\forall j = 2, \ldots, J$**.**

types. In particular, smaller (or non-state) banks have larger RTS estimates than their larger (or state) counterparts, respectively. This finding is in line with that in Altunbas, Liu, Molyneux and Seth (2000) who revealed the same inverse relationship between RTS and bank sizes using a sample of Japanese commercial banks from 1993 to 1996.

An advantage of the SPSC model is that it allows us to compute the marginal effects of the Z's on the distance elasticities with respect to the inputs and outputs;[25] therefore, we report in Table 3 the mean and quartile values (Q1–Q3) of the estimated marginal effects of Z_1 (i.e., deflated NPLs in log) and Z_2 (time) on the smooth coefficients from the constrained model that satisfies the theoretical properties of the IDF. The bootstrap and asymptotic standard errors are again reported in the parentheses and brackets, respectively, and are computed using the same methods as Table 2 does.[26] Most obviously, the time derivatives of the smooth coefficients allow us to compute TC, as shown in (23). In this equation, TC is decomposed into (1) a neutral component which captures the shift of the intercept of the IDF over time ($\partial\alpha/\partial t$) and (2) the two non-neutral components which capture shifts of the slopes of the IDF over time ($\sum_{k=1}^{K}\frac{\partial\gamma_k}{\partial t}\ln Y_k$ and $\sum_{j=2}^{J}\frac{\partial\beta_j}{\partial t}\ln\widetilde{X}_j$). More specifically, $\partial\beta/\partial Z_2$, where $Z_2 = t$, measures input bias in TC (Stevenson 1980). It quantifies reallocation of resources, i.e., changes in cost shares of inputs (β) over time, *ceteris paribus*. It can be seen that TC for most of the banks is found to be total-deposit-, fixed-assets-, and non-interest-expenses-using, given the significantly positive derivatives at the medians associated with β_1, β_2 and β_4, respectively; and interest-expenses-neutral, given the insignificant derivative at the median associated with β_3. These suggest that banks are trying to improve output quality by allocating more resources in credit and portfolio analysis. $\partial\gamma/\partial Z_2$, where $Z_2 = t$, measures scale bias in TC (Stevenson 1980). It quantifies changes in marginal costs of outputs (γ) over time, *ceteris paribus*. The medians of the derivatives of γ_1 and γ_3 are significantly positive, and this indicates that most banks are operating below their efficient scale for producing total loans (Y_1) and interest income (Y_3). However, most banks are

[25] We would like to thank an anonymous referee for pointing this out.

[26] It is worth noting that the bootstrap and asymptotic standard errors are the same to the fourth decimal place for all the mean values, but are quite different for some quartile values. Generally speaking, the asymptotic standard error of the mean does not require the quantity of interest to follow the normal (i.e., Gaussian) distribution, but those of the quartiles assume that the quantity of interest follows the normal distribution. Although the local-linear SPSC estimator is asymptotically normal (Li and Racine 2007, Geng and Sun 2019), the local-linear smooth coefficients and their derivative estimates from a finite sample may follow non-normal distributions. It turns out that the finite sample distributions of the marginal effect estimates are far away from the normal distribution—kernel density plots of the marginal effect estimates are omitted to save space, but are available upon request, and therefore the asymptotic standard errors of the quartiles would be misleading, and it is recommended that we use the bootstrap to compute the standard errors of the quartiles.

operating above their efficient scale for producing other earning assets (Y_2) and non-interest income (Y_4), given the significantly negative medians of the derivatives of γ_2 and γ_4. These findings reveal that bank mangers in Indonesia are risk-averse, and therefore they trade profit for reduced risk over time. Given that RTS depend on the marginal costs of outputs, we can calculate the marginal effects of the Z's on the RTS through their marginal effects on γ's.[27] It turns out that the marginal effect estimates of time on the RTS are heterogenous as they vary from -0.0092 to 0.0213. Generally, these marginal effects are negative from 2000 to 2008, and then become positive from 2009 to 2015 (i.e., after the global financial crisis). To consolidate the banking industry of Indonesia to be more resilient when confronted with any future financial crisis, mergers and acquisitions recently take place more often in Indonesia.

In addition to t, the other Z variable, i.e., Z_1 = deflated NPLs in log, can also shift the distance frontier in non-neutral manners, i.e., through the regression slopes, β's and γ's. The median $\partial\beta_1/\partial Z_1$ is significantly positive, and this means for most banks, an increase in NPLs would cause the cost share of total deposit to increase, *ceteris paribus*. The median derivatives of the other β_j's, $\forall j = 2, 3, 4$, with respect to Z_1 are all significantly negative, which means that for most banks, an increase in NPLs would cause the cost shares of the rest of the three inputs (i.e., fixed assets, interest expenses, and non-interest expenses) to decrease, *ceteris paribus*. These results indicate that when NPLs accumulate, depositors are paid a risk premium, and thus receive a higher interest rate as compensation for tolerating the extra loan risk (Hannan and Hanweck 1988, Hughes and Mester 1993); as banks become more susceptible to failure, the risk-averse bank managers become more reliant on deposit or short-term funding from interbank market to generate income. Furthermore, given that the median $\partial\gamma_1/\partial Z_1$ is significantly positive, i.e., $\partial(-\gamma_1)/\partial Z_1$ is significantly negative,[28] an increase in NPLs for most banks would cause the marginal cost of producing total loans (Y_1) to decrease, *ceteris paribus*. The median derivatives of the other γ_k's ($-\gamma_k$'s), $\forall k = 2, 3, 4$, with respect to Z_1 are all significantly negative (positive), which means that for most banks, an increase in NPLs would cause the marginal costs of producing all the other types of outputs (i.e., other earning assets, interest income, and non-interest income) to increase,

[27] **Since $RTS = -1/\sum_{k=1}^{K}\gamma_k(Z)$, then $\partial RTS/\partial Z_s = \left(\sum_{k=1}^{K}\gamma(Z)\right)^{-2}\cdot\sum_{k=1}^{K}\partial\gamma(Z)/\partial Z_s$, $\forall s = 1, 2$, where $s = 1$ for NPLs in log; $s = 2$ for time. Since $\left(\sum_{k=1}^{K}\gamma(Z)\right)^{-2}$ is positive, the sign of $\partial\ RTS/\partial\ Z_s$ depends on the sign of the sum of the derivatives of all the γ's with respect to a particular Z.**

[28] **Negation of the sign of γ's would cause the sign of the derivative of γ's to change, with the same bootstrap and asymptotic standard errors.**

***ceteris paribus*. This is because extra tasks, e.g., additional monitoring of the borrowers or renegotiating the terms on default loans, are demanded to resolve these NPLs. All these efforts would divert managers' attention away from normal daily monitoring of financial transactions, and this might impair banks' income-generating capabilities and increase marginal costs of producing income. In terms of the marginal effects of NPLs on RTS, it is found that the mean marginal effect is -0.0062 with a standard deviation of 0.0031. This means that, fix time, as NPLs increase by 1%, the RTS decreases by 0.0062/100=0.000062 units. Too many NPLs might cause banks' RTS to decrease because NPLs in most cases negatively affect banks' ability to transform their inputs into outputs.**[29]

A consistent estimation of all of the smooth coefficients is required to obtain consistent estimates of the non-linear stochastic frontier model, (17). Table 4 shows the MLE—η (slope vector for Z), σ_u^2, and σ_v^2—of the stochastic frontier model, and the least-squares cross-validated bandwidths for the Z's in the smooth coefficient model, (11). **To verify the advantage of the SPSC model over a traditional parametric model, the model specification test described in Section 2.4 is conducted where the null is the technology without Z (i.e., a parametric model of constant coefficients) and the alternative is the smooth-varying coefficient specification of technology that depends on Z. The bootstrapped p-value with 399 replications is zero to the fourth decimal place, and this means that the smooth coefficient model is preferred and the technology depends on the Z variables.**

In the SPSC stochastic frontier model, the technical inefficiency is conditionally heteroskedastic; that is, its pre-truncation variance is a function of Z—this is because variance better captures risk in production than mean. This specification of inefficiency allows us to obtain the exact (i.e., sign and magnitude of) marginal effects of Z on the inefficiency—(20) gives an example. However, we can easily know the direction of the marginal effects of a particular Z by looking at the sign of the coefficient associated with it. For example, the MLE for Z_1 in Table 4 is positive and significant at the 1% level. This indicates that an increase in NPLs would increase the pre-truncation variance of inefficiency, and this would also increase the post-truncation mean of inefficiency, and thus increase the inefficiency itself, holding the other Z variable constant. Similarly, there is strong evidence that inefficiency would decrease over time (Z_2), *ceteris paribus*. **To see a clearer picture of the marginal effects of NPLs and time on inefficiency, Figure 2 reports the histograms of these marginal effects, respectively. It can be seen that the marginal effects of NPLs on inefficiency are all positive, and the**

[29] **See Hughes and Mester (1998) who expressed scale economy as a function of NPLs.**

mean is 0.0207 with a standard deviation of 0.0122. Therefore, for an average bank, a 1% increase in NPLs would cause an input waste of 0.0207%, *ceteris paribus*. This result is unsurprising as extra costs are required to deal with the NPLs, e.g., additional credit screening and monitoring, disposal or sell-offs of the NPLs, etc.[30] On the contrary, the marginal effects of time on inefficiency are all negative, with a mean of -0.0122 and a standard deviation of 0.0073. This indicates that on average, 1.22% of the inputs are saved per annum—banks learn from past experiences for efficiency improvement. These results are consistent with those reported in Resti (1997), who found that under the stochastic frontier model, inefficiency decreased over time, and there was a strong negative correlation between NPLs and production efficiency for the Italian banking system during the period 1988 to 1992.

The TE scores are based on the MLE. Figure 3 reports the histogram of TE. The mean TE score is 0.83 with a standard deviation of 0.09. This mean TE score of banks is moderately high in Indonesia, which may partly be attributable to the recent market and regulatory reforms. It can be seen that the distribution of the estimated TE scores is skewed to the left. In fact, there are about 20% (80%) observations whose estimated TE's are greater (strictly less) than 0.9. This indicates that there is room for performance improvement for those banks that are not as efficient. To further investigate the TE scores, Figure 4 shows how the TE scores evolve over time. This is a box plot that shows the median, lower (i.e., 25th) and upper (i.e., 75th) quartiles, and inter-quartile ranges for the TE scores in each year. If the notches in the sides of the box plots do not overlap, then there is strong evidence that the two medians are statistically different from each other. We can see that: (1) the inter-quartile ranges as well as the differences between lower and upper quartiles generally fall over time, indicating that there is a slight decrease in variations of the TE scores over time; (2) the median TE scores generally increase over most years, indicating efficiency improvement of the Indonesian banking industry during most of the sample period. **This is in line with the MLE for Z_2 in Table 4, and with the histogram of the marginal effects of the time trend on inefficiency in Figure 2**, which indicates that inefficiency decreases over time; and (3) it is clear that the median TE scores from the year 2005 and onwards are significantly higher than that in 2000.

As a by-product of the SPSC input distance frontier model, TFP growth and its components are computed using the estimated smooth coefficients and inefficiencies, along with their time derivatives. Figure 5 plots the weighted average of TFP growth and its components over time where the weights

[30] **Mester (1996) and Berger and DeYoung (1997) explained the reasons for controlling for NPLs when estimating banks' level of inefficiency.**

are determined by the sizes (i.e., log of total assets) of banks. These growth rates are used to define their respective indices, as reported in Figure 6, from $X_t = X_{t-1}(1 + \dot{X}_t)$, where X is either TFP or one of its components, and $X_{2000} = 100$. A positive (negative) growth rate in year t indicates that the corresponding index would rise (fall) from year $t-1$ to year t. Thus, these indices reveal the temporal behavior of TFP growth and its components.

We can see from Figure 5 that (1) the EC and TC components almost cancel out each other for most of the sample period; (2) the resultant TFP growth (i.e., the sum of TC, SC, and EC) generally follows the trajectory of the SC without any surprise; and (3) the Divisia measure of TFP growth is more fluctuative since it includes noises from data while the estimated TFP growth from the SPSC model has accommodated these noises via regression. In Figure 6, unsurprisingly, the TFP growth index generally remains stable over the sample period. That the EC and TC components move in opposite directions, the SC most closely tracks the temporal behavior of TFP growth, and the Divisia index is more fluctuative confirm the results reported in Figure 5. These two figures imply that the EC is the only source of TFP growth in the Indonesian banking industry, and efforts should be made in enhancing technical progress and increasing the effect of scale economy, and thus effectively boosting TFP growth and index. These might be achieved through further reforms and policies that could mitigate potentially upward pressure on market concentration. Necessary steps must be taken to modernize legal and political institutions, comply with international standards such as following the Basel Accords, and maintain macroeconomic stability to realize the potential of the sector.

5 Conclusion

This paper proposes a semiparametric smooth-varying coefficient panel stochastic input distance frontier with determinants of inefficiency, where the technology parameters are unknown functions of a set of environmental factors that non-neutrally shift the distance frontier **in a fully flexible manner**. Homogeneity and monotonicity constraints are imposed on the distance function via the computationally simple CWB method. Furthermore, TFP growth is decomposed into TC, SC, and EC. We apply the methodology to an unbalanced panel data on the Indonesian banking industry during the period 2000–2015 with multiple outputs and multiple inputs, and estimate the distance function elasticities, marginal effects of the environmental factors on the elasticities and inefficiency, RTS, TE scores, and TFP growth and its components. We find that (1) on average, total deposit has the largest share of cost in producing the outputs; (2) it is most costly to produce an additional percent of interest income; (3) generally speaking, NPLs have adverse effect on the banking sector since they increase the marginal

costs of producing most outputs and cause the inefficiency to increase, *ceteris paribus*; (4) the median TE scores generally increase over the years; and (5) TFP growth generally follows the trajectory of the SC. Our policy implication is that efforts should be made to enhance technological progress and increasing the effect of scale economy so as to promote TFP growth. This research can be replicated to investigate banks' technology, including their TE and TFP growth patterns, for other developing economies.

Declarations

- Conflict of Interest: The authors declare that they have no conflict of interest.

References

Aigner, D. J., Lovell, C. A. K. and Schmidt, P. (1977), 'Formulation and estimation of stochastic frontier production functions', *Journal of Econometrics* **6**(1), 21–37.

Altunbas, Y., Liu, M.-H., Molyneux, P. and Seth, R. (2000), 'Efficiency and risk in Japanese banking', *Journal of Banking and Finance* **24**(10), 1605–1628.

Assaf, A. G., Matousek, R. and Tsionas, E. G. (2013), 'Turkish bank efficiency: Bayesian estimation with undesirable outputs', *Journal of Banking and Finance* **37**(2), 506–517.

Bank Indonesia (2012–2017), Indonesian banking statistics, Report.

Berger, A. N. and DeYoung, R. (1997), 'Problem loans and cost efficiency in commercial banks', *Journal of Banking and Finance* **21**(6), 849–870.

Berger, A. N. and Humphrey, D. B. (1997), 'Efficiency of financial institutions: International survey and directions for future research', *European Journal of Operational Research* **98**(2), 175–212.

Bhaumik, S. K., Das, P. K. and Kumbhakar, S. C. (2012), 'A stochastic frontier approach to modelling financial constraints in firms: An application to India', *Journal of Banking and Finance* **36**(5), 1311–1319.

Bhaumik, S. K., Kumbhakar, S. C. and Sun, K. (2015), 'A note on a semiparametric approach to estimating financing constraints in firms', *European Journal of Finance* **21**, 992–1004.

Cai, Z., Fan, J. and Li, R. (2000), 'Efficient estimation and inferences for varying-coefficient models', *Journal of the American Statistical Association* **95**(451), 888–902.

Cai, Z., Fan, J. and Yao, Q. (2000), 'Functional-coefficient regression models for nonlinear time series', *Journal of the American Statistical Association* **95**(451), 941–956.

Cameron, A. and Trivedi, K. (2005), *Microeconometrics: Methods and Application*, Cambridge University Press, chapter 11, pp. 376–377.

Chen, R. and Tsay, R. (1993), 'Functional-coefficient autoregressive models', *Journal of the American Statistical Association* **88**, 298–308.

Das, A. and Kumbhakar, S. C. (2012), 'Productivity and efficiency dynamics in Indian banking: An input distance function approach incorporating quality of inputs and outputs', *Journal of Applied Econometrics* **27**(2), 205–234.

Defung, F., Salim, R. and Bloch, H. (2016), 'Has regulatory reform had any impact on bank efficiency in Indonesia? A two-stage analysis', *Applied Economics* **48**, 5060–5074.

Denizer, C. (2000), Foreign entry in Turkey's banking sector, 1980–97, Policy Research Working Paper Series 2462, The World Bank, Washington D.C.

Diewert, W. E. and Wales, T. J. (1987), 'Flexible functional forms and global curvature conditions', *Econometrica* **55**(1), 43–68.

Du, P., Parmeter, C. and Racine, J. (2013), 'Nonparametric kernel regression with multiple predictors and multiple shape constraints', *Statistica Sinica* **23**(3), 1347–1371.

Esho, N. (2001), 'The determinants of cost efficiency in cooperative financial institutions: Australian evidence', *Journal of Banking and Finance* **25**(5), 941–964.

Fan, J. and Zhang, W. (1999), 'Statistical estimation in varying-coefficient models', *The Annals of Statistics* **27**, 1491–1518.

Fan, Y. and Li, Q. (1996), 'Consistent model specification tests: Omitted variables and semiparametric functional forms', *Econometrica* **64**, 865–890.

Feng, G. and Serletis, A. (2010), 'Efficiency, technical change, and returns to scale in large US banks: Panel data evidence from an output distance function satisfying theoretical regularity', *Journal of Banking and Finance* **34**(1), 127–138.

Fofack, H. L. (2005), Nonperforming loans in Sub-Saharan Africa: Causal analysis and macroeconomic implications, Policy Research Working Paper Series 3769, The World Bank, Washington D.C.

Fries, S. and Taci, A. (2005), 'Cost efficiency of banks in transition: Evidence from 289 banks in 15 post-communist countries', *Journal of Banking and Finance* **29**(1), 55–81.

Geng, X. and Sun, K. (2019), 'Gradient estimation of the local-constant semiparametric smooth coefficient model', *Economics Letters* **185**, 108684.

Gunes, H. and Yildirim, D. (2016), 'Estimating cost efficiency of Turkish commercial banks under unobserved heterogeneity with stochastic frontier models', *Central Bank Review* **16**(4), 127–136.

Hadad, M. D., Hall, M. J. B., Santoso, W., Satria, R., Kenjegalieva, K. and Simper, R. (2008), Efficiency in Indonesian banking: Recent evidence, Discussion paper series, Department of Economics, Loughborough University.

Hall, P. and Huang, H. (2001), 'Nonparametric kernel regression subject to monotonicity constraints', *The Annals of Statistics* **29**(3), 624–647.

Hannan, T. and Hanweck, G. (1988), 'Bank insolvency risk and the market for large certificates of deposit', *Journal of Money, Credit and Banking* **20**(2), 203–211.

Hastie, T. and Tibshirani, R. (1993), 'Varying-coefficient models', *Journal of the Royal Statistical Society. Series B (Methodological)* **55**(4), 757–796.

Heshmati, A., Kumbhakar, S. C. and Sun, K. (2014), 'Estimation of productivity in Korean electric power plants: A semiparametric smooth coefficient model', *Energy Economics* **45**, 491–500.

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', *The Review of Economics and Statistics* **80**(2), 314–325.

Koenker, R. W. and Bassett, G. W. (1978), 'Regression quantiles', *Econometrica* **46**, 33–50.

Kumbhakar, S. C. (2013), 'Specification and estimation of multiple output technologies: A primal approach', *European Journal of Operational Research* **231**, 465–473.

Kumbhakar, S. C., Park, B. U., Simar, L. and Tsionas, E. G. (2007), 'Nonparametric stochastic frontiers: A local maximum likelihood approach', *Journal of Econometrics* **137**(1), 1–27.

Kumbhakar, S. C. and Sun, K. (2012), 'Estimation of TFP growth: A semiparametric smooth coefficient approach', *Empirical Economics* **43**, 1–24.

Kumbhakar, S. C. and Wang, H.-J. (2005), 'Estimation of growth convergence using a stochastic production frontier approach', *Economics Letters* **88**(3), 300–305.

Li, Q., Huang, C., Li, D. and Fu, T. (2002), 'Semiparametric smooth coefficient models', *Journal of Business and Economic Statistics* **20**(3), 412–422.

Li, Q. and Racine, J. (2007), *Nonparametric Econometrics: Theory and Practice*, Princeton University Press.

Li, Q. and Racine, J. (2010), 'Smooth varying-coefficient estimation and inference for qualitative and quantitative data', *Econometric Theory* **26**, 1–31.

Longford, N. T. (1994), *Random Coefficient Models (Oxford Statistical Science Series) 1st Edition*, Clarendon Press.

Meeusen, W. and van den Broeck, J. (1977), 'Efficiency estimation from Cobb-Douglas production functions with composed error', *International Economic Review* **18**(2), 435–444.

Mester, L. J. (1996), 'A study of bank efficiency taking into account risk-preferences', *Journal of Banking and Finance* **20**(6), 1025–1045.

O'Donnell, C. J. and Coelli, T. J. (2005), 'A Bayesian approach to imposing curvature on distance functions', *Journal of Econometrics* **126**(2), 493–523.

O'Donnell, C. J., Rambaldi, A. N. and Doran, H. E. (2001), 'Estimating economic relationships subject to firm- and time-varying equality and inequality constraints', *Journal of Applied Econometrics* **16**(4), 709–726.

Park, K. H. and Weber, W. L. (2006), 'A note on efficiency and productivity growth in the Korean banking industry, 1992–2002', *Journal of Banking and Finance* **30**(8), 2371–2386.

Parmeter, C. F., Sun, K., Henderson, D. J. and Kumbhakar, S. C. (2014), 'Estimation and inference under economic restrictions', *Journal of Productivity Analysis* **41**(1), 111–129.

Podpiera, J. and Weill, L. (2008), 'Bad luck or bad management? Emerging banking market experience', *Journal of Financial Stability* **4**(2), 135–148.

Resti, A. (1997), 'Evaluating the cost-efficiency of the Italian banking system: What can be learned from the joint application of parametric and non-parametric techniques', *Journal of Banking and Finance* **21**(2), 221–250.

Sealey, C. W. and Lindley, J. T. (1977), 'Inputs, outputs, and a theory of production and cost at depository financial institutions', *The Journal of Finance* **32**(4), 1251–1266.

Servin, R., Lensink, R. and van den Berg, M. (2012), 'Ownership and technical efficiency of microfinance institutions: Empirical evidence from Latin America', *Journal of Banking and Finance* **36**(7), 2136–2144.

Shephard, R. (1953), *Cost and Production Functions*, Princeton University Press, Princeton.

Shephard, R. (1970), *Theory of Cost and Production Functions*, Princeton University Press, Princeton.

Stevenson, R. (1980), 'Measuring technological bias', *American Economic Review* **70**, 162–173.

Sun, K. (2015), 'Constrained nonparametric estimation of the input distance function', *Journal of Productivity Analysis* **43**(1), 85–97.

Sun, K. and Kumbhakar, S. C. (2013), 'Semiparametric smooth-coefficient stochastic frontier model', *Economic Letters* **120**(2), 305–309.

Sun, K., Kumbhakar, S. C. and Tveterås, R. (2015), 'Productivity and efficiency estimation: A semiparametric stochastic cost frontier approach', *European Journal of Operational Research* **245**(1), 194–202.

Terrell, D. (1996), 'Incorporating monotonicity and concavity conditions in flexible functional forms', *Journal of Applied Econometrics* **11**(2), 179–194.

Wang, H.-J. and Ho, C.-W. (2010), 'Estimating fixed-effect panel stochastic frontier models by model transformation', *Journal of Econometrics* **157**, 286–296.

Wang, H.-J. and Schmidt, P. (2002), 'One-step and two-step estimation of the effects of exogenous variables on technical efficiency levels', *Journal of Productivity Analysis* **18**, 129–144.

Yao, F., Zhang, F. and Kumbhakar, S. C. (2019), 'Semiparametric smooth coefficient stochastic frontier model with panel data', *Journal of Business & Economic Statistics* **37**(3), 556–572.

Table 1: **Summary Statistics of the Variables**

Symbol	Variable Name	Mean	Sd.	Min.	Max.
Inputs					
X_1	Total deposit	3529939.45	10953662.21	1016.98	127845375.30
X_2	Fixed assets	100370.99	324979.85	396.58	5011884.79
X_3	Interest expenses	184226.02	558291.96	63.15	10357553.72
X_4	Non-interest expenses	191394.75	582949.56	307.93	7518004.83
Outputs					
Y_1	Total loans	2558936.70	8252872.76	377.00	109135131.30
Y_2	Other earning assets	1295597.83	3991240.21	1931.25	51017045.59
Y_3	Interest income	408601.34	1208441.26	314.72	16068145.81
Y_4	Non-interest income	89884.06	335931.97	10.25	5774588.57
Environmental factors					
Z_1	Deflated NPLs in log	9.0882	2.3632	-0.2722	15.0663
Z_2	Time trend (t)	8.4354	4.6002	1	16

1. Total number of observations = 1493.
2. t is calculated as year - 1999, where year goes from 2000 to 2015.

Figure 1: **Kernel Density Plots of Estimated Smooth Coefficients**

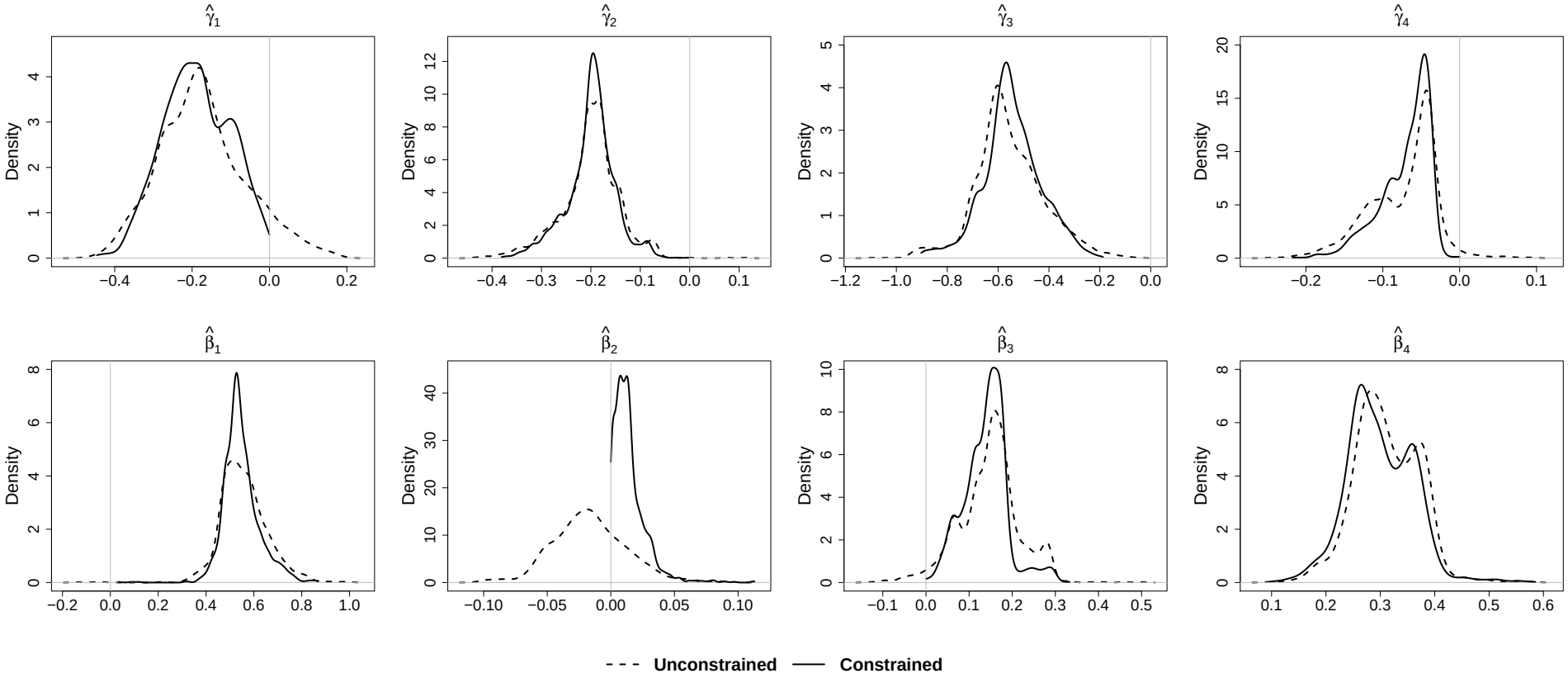


Table 2: **Summary Statistics of the Smooth Coefficients**

Model	γ_1	γ_2	γ_3	γ_4	β_1	β_2	β_3	β_4
Unconstrained								
Mean	-0.1729	-0.2009	-0.5547	-0.0734	0.5575	-0.0164	0.1475	0.3114
	(0.0029) [0.0029]	(0.0016) [0.0015]	(0.0034) [0.0035]	(0.0011) [0.0011]	(0.0025) [0.0026]	(0.0008) [0.0007]	(0.0018) [0.0018]	(0.0015) [0.0015]
Q1	-0.2517	-0.2279	-0.6322	-0.1041	0.4933	-0.0350	0.1107	0.2711
	(0.0041) [0.0043]	(0.0026) [0.0024]	(0.0038) [0.0034]	(0.0025) [0.0023]	(0.0021) [0.0022]	(0.0009) [0.0010]	(0.0025) [0.0024]	(0.0014) [0.0014]
Q2	-0.1802	-0.1960	-0.5726	-0.0609	0.5473	-0.0177	0.1533	0.3056
	(0.0029) [0.0028]	(0.0017) [0.0017]	(0.0032) [0.0034]	(0.0021) [0.0017]	(0.0034) [0.0033]	(0.0008) [0.0008]	(0.0010) [0.0010]	(0.0026) [0.0021]
Q3	-0.1089	-0.1699	-0.4783	-0.0417	0.6130	0.0014	0.1841	0.3577
	(0.0053) [0.0047]	(0.0020) [0.0020]	(0.0037) [0.0038]	(0.0004) [0.0004]	(0.0034) [0.0034]	(0.0011) [0.0010]	(0.0016) [0.0015]	(0.0026) [0.0027]
% of violations	7.77%	0.13%	0.00%	1.94%	0.20%	73.34%	2.61%	0.00%
Constrained								
Mean	-0.1876	-0.1994	-0.5448	-0.0703	0.5472	0.0143	0.1391	0.2994
	(0.0022) [0.0023]	(0.0013) [0.0013]	(0.0029) [0.0030]	(0.0008) [0.0008]	(0.0019) [0.0020]	(0.0003) [0.0003]	(0.0013) [0.0013]	(0.0016) [0.0016]
Q1	-0.2495	-0.2227	-0.6059	-0.0887	0.5028	0.0060	0.1091	0.2585
	(0.0029) [0.0029]	(0.0021) [0.0021]	(0.0020) [0.0021]	(0.0015) [0.0015]	(0.0023) [0.0022]	(0.0002) [0.0002]	(0.0019) [0.0019]	(0.0010) [0.0010]
Q2	-0.1909	-0.1961	-0.5534	-0.0613	0.5349	0.0119	0.1445	0.2930
	(0.0042) [0.0040]	(0.0011) [0.0011]	(0.0026) [0.0026]	(0.0014) [0.0015]	(0.0016) [0.0015]	(0.0003) [0.0003]	(0.0016) [0.0014]	(0.0021) [0.0020]
Q3	-0.1193	-0.1729	-0.4781	-0.0456	0.5811	0.0187	0.1695	0.3468
	(0.0037) [0.0038]	(0.0017) [0.0015]	(0.0046) [0.0044]	(0.0005) [0.0005]	(0.0033) [0.0031]	(0.0006) [0.0006]	(0.0012) [0.0011]	(0.0024) [0.0024]
p-value	0.0000	0.0016	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

1. β_1 is calculated as $1 - \sum_{j=2}^{J} \beta_j$, for the unconstrained and constrained models.
2. Bootstrapped standard errors are reported in the parentheses, and asymptotic standard errors are reported in the brackets.
3. Percentages of violations are reported under the unconstrained model.
4. p-values from the Kolmogorov-Smirnov test comparing the unconstrained and constrained models are reported under the constrained model.

Table 3: **Summary Statistics of the Marginal Effects from the Constrained Model**

Z_1 (NPLs)	$\frac{\partial \gamma_1}{\partial Z_1}$	$\frac{\partial \gamma_2}{\partial Z_1}$	$\frac{\partial \gamma_3}{\partial Z_1}$	$\frac{\partial \gamma_4}{\partial Z_1}$	$\frac{\partial \beta_1}{\partial Z_1}$	$\frac{\partial \beta_2}{\partial Z_1}$	$\frac{\partial \beta_3}{\partial Z_1}$	$\frac{\partial \beta_4}{\partial Z_1}$
Mean	0.0150	-0.0095	-0.0100	-0.0017	0.0268	-0.0027	-0.0100	-0.0142
	(0.0005)	(0.0004)	(0.0008)	(0.0002)	(0.0005)	(0.0001)	(0.0002)	(0.0005)
	[0.0005]	[0.0004]	[0.0008]	[0.0002]	[0.0005]	[0.0001]	[0.0002]	[0.0005]
Q1	-0.0027	-0.0216	-0.0351	-0.0078	0.0084	-0.0043	-0.0168	-0.0293
	(0.0019)	(0.0005)	(0.0004)	(0.0007)	(0.0021)	(0.0001)	(0.0001)	(0.0010)
	[0.0015]	[0.0003]	[0.0004]	[0.0003]	[0.0003]	[0.0001]	[0.0001]	[0.0002]
Q2	0.0116	-0.0075	-0.0249	-0.0018	0.0242	-0.0027	-0.0102	-0.0123
	(0.0051)	(0.0009)	(0.0012)	(0.0001)	(0.0007)	(0.0001)	(0.0003)	(0.0021)
	[0.0008]	[0.0003]	[0.0010]	[0.0001]	[0.0006]	[0.0001]	[0.0003]	[0.0003]
Q3	0.0339	0.0050	0.0204	0.0012	0.0416	-0.0010	-0.0051	0.0021
	(0.0003)	(0.0012)	(0.0011)	(0.0008)	(0.0008)	(0.0001)	(0.0002)	(0.0023)
	[0.0003]	[0.0009]	[0.0012]	[0.0006]	[0.0008]	[0.0000]	[0.0002]	[0.0006]
Z_2 (t)	$\frac{\partial \gamma_1}{\partial Z_2}$	$\frac{\partial \gamma_2}{\partial Z_2}$	$\frac{\partial \gamma_3}{\partial Z_2}$	$\frac{\partial \gamma_4}{\partial Z_2}$	$\frac{\partial \beta_1}{\partial Z_2}$	$\frac{\partial \beta_2}{\partial Z_2}$	$\frac{\partial \beta_3}{\partial Z_2}$	$\frac{\partial \beta_4}{\partial Z_2}$
Mean	-0.0140	-0.0098	0.0293	-0.0047	0.0079	0.0015	-0.0164	0.0070
	(0.0011)	(0.0008)	(0.0019)	(0.0003)	(0.0006)	(0.0001)	(0.0008)	(0.0006)
	[0.0011]	[0.0008]	[0.0019]	[0.0003]	[0.0006]	[0.0001]	[0.0008]	[0.0006]
Q1	-0.0592	-0.0440	-0.0018	-0.0146	-0.0054	-0.0002	-0.0433	-0.0158
	(0.0137)	(0.0128)	(0.0036)	(0.0010)	(0.0023)	(0.0001)	(0.0007)	(0.0003)
	[0.0018]	[0.0010]	[0.0008]	[0.0001]	[0.0001]	[0.0001]	[0.0007]	[0.0003]
Q2	0.0006	-0.0024	0.0100	-0.0032	0.0137	0.0014	-0.0032	0.0162
	(0.0003)	(0.0005)	(0.0008)	(0.0003)	(0.0007)	(0.0002)	(0.0023)	(0.0008)
	[0.0003]	[0.0005]	[0.0004]	[0.0002]	[0.0005]	[0.0001]	[0.0013]	[0.0004]
Q3	0.0125	0.0151	0.1104	0.0079	0.0226	0.0056	0.0086	0.0278
	(0.0002)	(0.0007)	(0.0380)	(0.0021)	(0.0007)	(0.0002)	(0.0001)	(0.0005)
	[0.0002]	[0.0003]	[0.0030]	[0.0000]	[0.0007]	[0.0002]	[0.0001]	[0.0004]

1. $\partial \beta_1 / \partial Z_s$ are calculated as $-\sum_{j=2}^{J} \partial \beta_j / \partial Z_s$, $\forall s = 1, 2$.
2. Bootstrapped standard errors are reported in the parentheses, and asymptotic standard errors are reported in the brackets.

Table 4: **Maximum Likelihood Estimates and Bandwidths**

	MLE	SE	Bandwidth
Z_1 (NPLs)	0.1042	0.0288	6.5868
Z_2 (t)	-0.0618	0.0101	1.5867
σ_u^2	0.0225	0.0124	-
σ_v^2	0.0204	0.0008	-

The first column gives the maximum likelihood estimates of the non-linear stochastic frontier model as in (17). The second column gives the corresponding standard error for each estimate. The third column reports the least-squares cross-validated bandwidths for the Z variables.

Figure 2: **Marginal Effects of Z on Inefficiency**

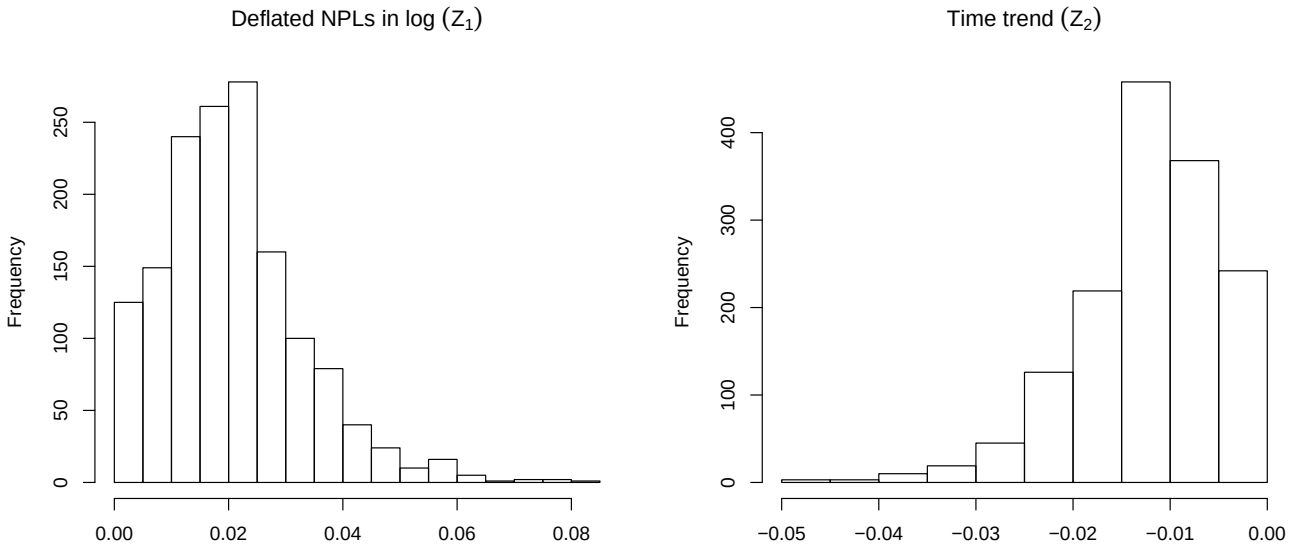


Figure 3: **TE Scores**

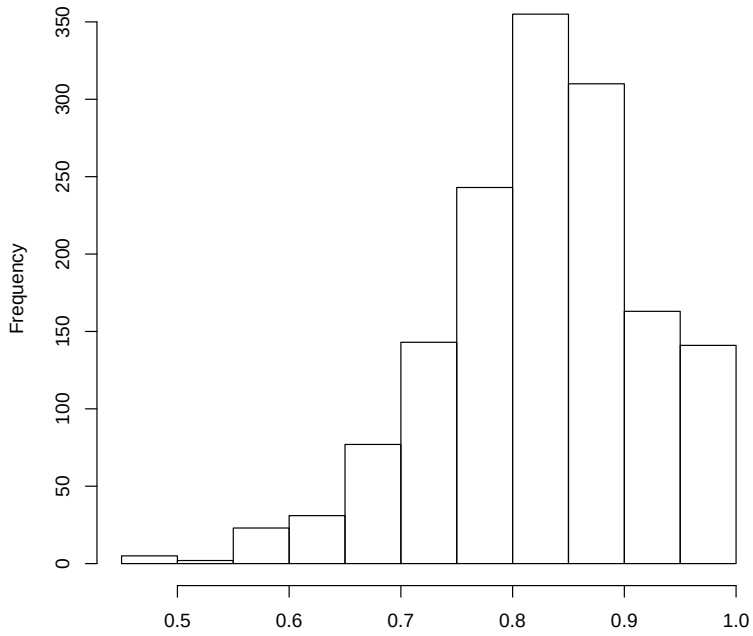


Figure 4: **TE Scores Over Time**

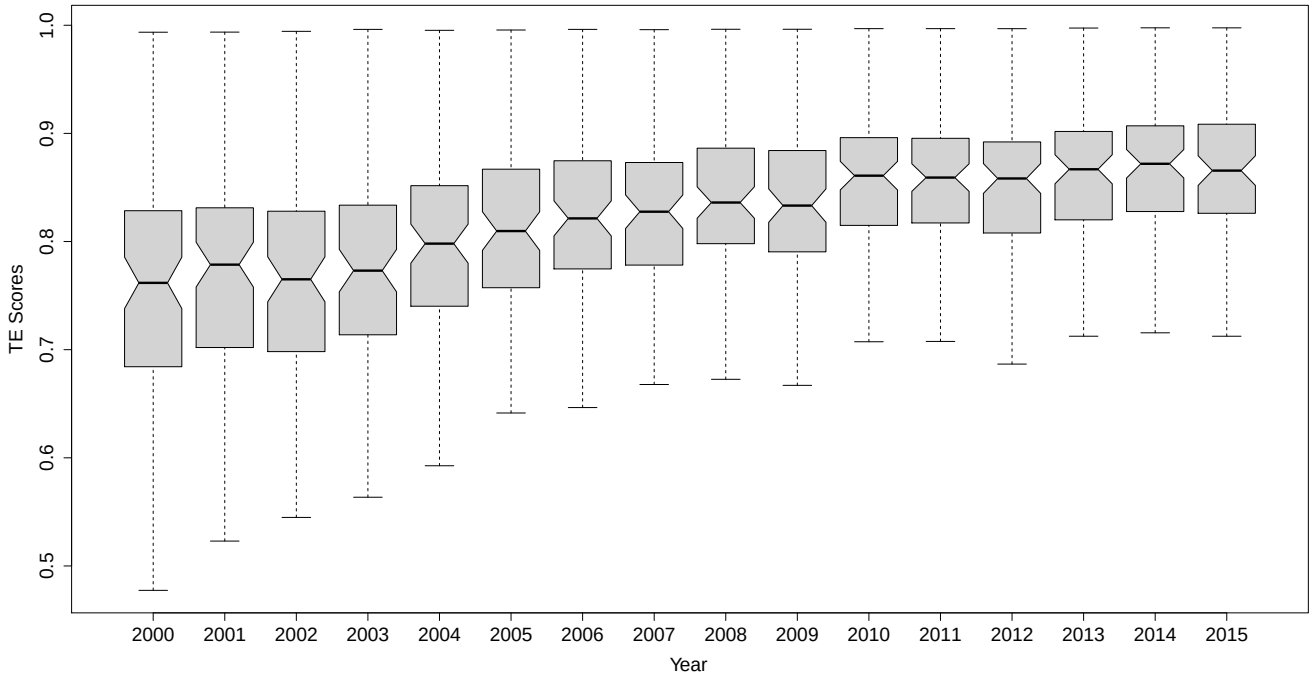


Figure 5: **TFP Growth and Its Components**

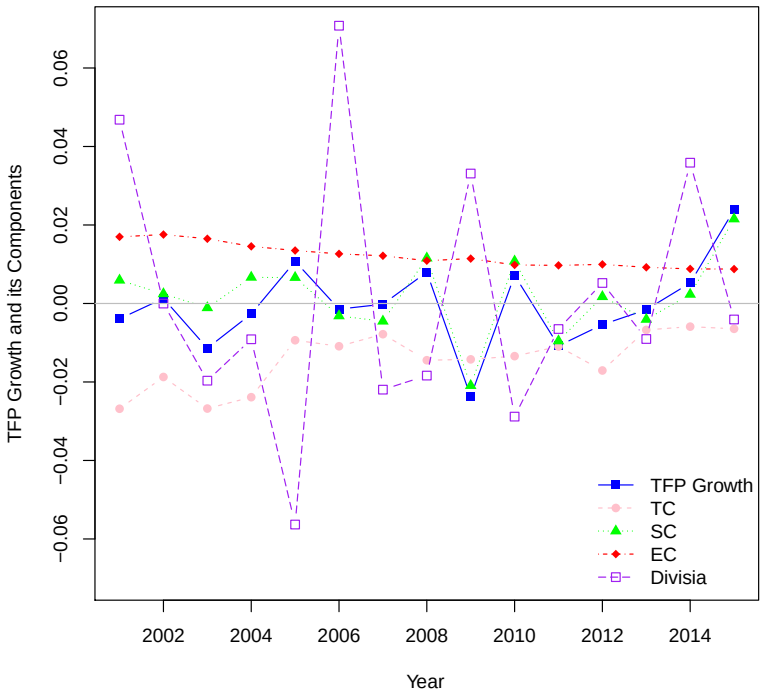
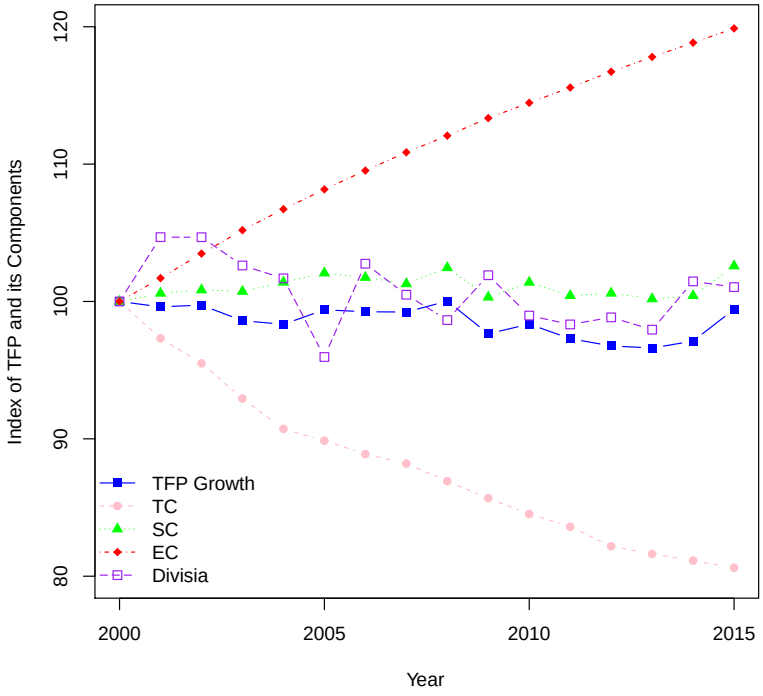


Figure 6: **TFP Index and Its Components**



Appendix A

This appendix explains how we select the bandwidths to estimate the smooth coefficients and their derivatives in (11). Following Li and Racine (2010), we employ the most commonly used least-squares cross-validation (LSCV) method, which is a fully automatic data-driven approach, to select the bandwidth vector h; that is,

$$CV_{ll}(h) = \min_{h} \sum_{i=1}^{N} \sum_{t=1}^{T} [\mathcal{Y}_{it} - W'_{it}\hat{\rho}_{-it}(Z_{it})]^2 M(Z_{it}), \tag{33}$$

where $CV_{ll}(h)$ determines the cross-validation bandwidth vector h for local-linear estimator, $W'_{it}\hat{\rho}_{-it}(Z_{it})$ is the leave-one-out local-linear kernel conditional mean, and $0 \leq M(\cdot) \leq 1$ is a weight function that serves to avoid difficulties caused by dividing by zero. The same bandwidth vector is used to estimate the constrained smooth coefficients.

Appendix B

This appendix details the TFP growth decomposition given in Section 2.3.

First, take the time derivatives of both sides of (4) (ignoring the noise term), and we would have:

$$-\dot{X}_1 = \frac{\partial \alpha(Z)}{\partial t} + \sum_{k=1}^{K} \frac{\partial \gamma_k(Z)}{\partial t} \ln Y_k + \sum_{k=1}^{K} \gamma_k(Z)\dot{Y}_k + \sum_{j=2}^{J} \frac{\partial \beta_j(Z)}{\partial t} \ln \widetilde{X}_j + \sum_{j=2}^{J} \beta_j(Z)\dot{\widetilde{X}}_j - \frac{u}{t}. \tag{34}$$

The TC component measures the non-neutral shift of the IDF over time; therefore,

$$-\dot{X}_1 = TC + \sum_{k=1}^{K} \gamma_k(Z)\dot{Y}_k + \sum_{j=2}^{J} \beta_j(Z)\dot{\widetilde{X}}_j + EC, \tag{35}$$

where $TC = \frac{\partial \alpha(Z)}{\partial t} + \sum_{k=1}^{K} \frac{\partial \gamma_k(Z)}{\partial t} \ln Y_k + \sum_{j=2}^{J} \frac{\partial \beta_j(Z)}{\partial t} \ln \widetilde{X}_j$, and $EC = -\frac{u}{t}$ measures the efficiency change.

Add $T\dot{F}P$ to both sides of (35) and rearrange to obtain:

$$\begin{aligned} T\dot{F}P &= TC + T\dot{F}P + \dot{X}_1 + \sum_{k=1}^{K} \gamma_k(Z)\dot{Y}_k + \sum_{j=2}^{J} \beta_j(Z)\dot{\widetilde{X}}_j + EC \\ &= TC + \sum_{k=1}^{K} R_k\dot{Y}_k - \sum_{j=1}^{J} S_j\dot{X}_j + \dot{X}_1 + \sum_{k=1}^{K} \gamma_k(Z)\dot{Y}_k + \sum_{j=2}^{J} \beta_j(Z)\dot{\widetilde{X}}_j + EC \\ &= TC + SC + \dot{X}_1 - \sum_{j=1}^{J} S_j\dot{X}_j + \sum_{j=2}^{J} \beta_j(Z)\dot{\widetilde{X}}_j + EC, \end{aligned} \tag{36}$$

where $SC = \sum_{k=1}^{K} R_k \dot{Y}_k + \sum_{k=1}^{K} \gamma_k(Z)\dot{Y}_k = \sum_{k=1}^{K}(R_k + \gamma_k(Z))\dot{Y}_k$. To further decompose the SC into one sub-component related to RTS and the other related to market power:

$$\begin{aligned} SC &= \sum_{k=1}^{K} R_k \dot{Y}_k + \sum_{k=1}^{K} \gamma_k(Z)\dot{Y}_k + RTS \cdot \sum_{k=1}^{K} \gamma_k(Z)\dot{Y}_k - RTS \cdot \sum_{k=1}^{K} \gamma_k(Z)\dot{Y}_k \\ &= (1 - RTS)\sum_{k=1}^{K} \gamma_k(Z)\dot{Y}_k + \sum_{k=1}^{K} R_k \dot{Y}_k - \sum_{k=1}^{K} \frac{\gamma_k(Z)}{\sum_{k=1}^{K}\gamma_k(Z)}\dot{Y}_k \\ &= (1 - RTS)\sum_{k=1}^{K} \gamma_k(Z)\dot{Y}_k + \sum_{k=1}^{K}\left(R_k - \frac{\gamma_k(Z)}{\sum_{k=1}^{K}\gamma_k(Z)}\right)\dot{Y}_k, \end{aligned} \tag{37}$$

given that $RTS = -1/\sum_{k=1}^{K}\gamma_k(Z)$. Finally, it can be shown that:

$$\begin{aligned} AC &= \dot{X}_1 - \sum_{j=1}^{J} S_j \dot{X}_j + \sum_{j=2}^{J} \beta_j(Z)\dot{\tilde{X}}_j \\ &= \sum_{j=1}^{J} S_j \dot{X}_1 - \sum_{j=1}^{J} S_j \dot{X}_j + \sum_{j=2}^{J} \beta_j(Z)\dot{\tilde{X}}_j \\ &= \sum_{j=1}^{J}(\dot{X}_1 - \dot{X}_j)S_j + \sum_{j=2}^{J} \beta_j(Z)\dot{\tilde{X}}_j, \end{aligned} \tag{38}$$

using the fact that $\sum_{j=1}^{J} S_j = 1$. Furthermore,

$$\begin{aligned} \sum_{j=1}^{J}(\dot{X}_1 - \dot{X}_j)S_j = -\sum_{j=1}^{J}(\dot{X}_j - \dot{X}_1)S_j &= -\sum_{j=1}^{J}\dot{\tilde{X}}_j S_j \\ &= -\left(\dot{\tilde{X}}_1 S_1 + \sum_{j=2}^{J}\dot{\tilde{X}}_j S_j\right) = -\sum_{j=2}^{J}\dot{\tilde{X}}_j S_j, \end{aligned} \tag{39}$$

using the fact that $\dot{\tilde{X}}_1 = 0$. Therefore, plug (39) into (38) to obtain $AC = \sum_{j=2}^{J}(\beta_j(Z) - S_j)\dot{\tilde{X}}_j$. QED.