Empir Econ (2015) 49:185–211
DOI 10.1007/s00181-014-0849-z

Estimation of banking technology under credit uncertainty

Emir Malikov · Diego Restrepo-Tobón · Subal C. Kumbhakar

Received: 19 November 2013 / Accepted: 15 May 2014 / Published online: 27 July 2014

Abstract Credit risk is crucial to understanding banks' production technology and should be explicitly accounted for when modeling the latter. The banking literature has largely accounted for risk using *ex-post* realizations of banks' uncertain outputs and the variables intended to capture risk. This is equivalent to estimating an *ex-post* realization of bank's production technology which, however, may not reflect optimality conditions that banks seek to satisfy under uncertainty. The *ex-post* estimates of technology are likely to be biased and inconsistent, and one thus may call into question the reliability of the results regarding banks' technological characteristics broadly reported in the literature. However, the extent to which these concerns are relevant for policy analysis is an empirical question. In this paper, we offer an alternative methodology to estimate banks' production technology based on the *ex-ante* cost function. We model credit uncertainty explicitly by recognizing that bank managers minimize costs subject to given *expected* outputs and credit risk. We estimate unobservable expected outputs and associated credit risk levels from banks' supply functions via nonparametric kernel methods. We apply this framework to estimate production technology of U.S. commercial banks during the period from 2001 to 2010 and contrast the new estimates with those based on the *ex-post* models widely employed in the literature.

E. Malikov (✉)
Department of Economics, State University of New York at Binghamton, P.O. Box 6000, Binghamton, NY 13902, USA
e-mail: emalikov@binghamton.edu

E. Malikov
Department of Economics, St. Lawrence University, Canton, NY, USA

D. Restrepo-Tobón
Department of Finance, EAFIT University, Medellín, Colombia

S. C. Kumbhakar
Department of Economics, State University of New York at Binghamton, Binghamton, NY, USA

Keywords Ex-ante cost function · Production uncertainty · Productivity · Returns to scale · Risk

JEL Classification C10 · D81 · G21

1 Introduction

Commercial banking is a risky business. Banks are subject to risks of many kinds among which credit risk and liquidity risk are the two most commonly referenced (e.g., Freixas and Rochet 2008). Credit risk is associated with the likelihood that a borrower will default on the debt by failing to make payments as obligated contractually. Liquidity risk arises from financing long-term illiquid assets with short-term liquid liabilities. These risks shape banks' production technology by requiring them to spend substantial resources on risk management. Therefore, researchers should explicitly account for these intrinsic risks when modeling banks' production technology (Hughes and Mester 1998).

A handful of empirical studies on microeconomics of banking account for risk-taking behavior of banks. With a few exceptions (Hughes et al. 1996; Hughes et al. 2000, 2001; Hughes and Mester 2013), researchers customarily estimate banks' scale and scope economies, productivity or efficiency by assuming away uncertainty in banks' production (e.g., Berger et al. 1987; Clark 1996; Berger and Mester 1997, 2003; Wheelock and Wilson 2001, 2012; Feng and Serletis 2009, 2010). Credit risk is often (if at all) modeled using its *ex-post* "proxies" such as non-performing loans or the volatility of net income in prior years (Hughes and Mester 1993), while liquidity risk is seldom accounted for.[1]

The use of *ex-post* realizations of risk and uncertain outputs (to which all above-cited papers resort to) is equivalent to estimating an *ex-post* realization of banks' production technology which, however, may not reflect actual optimality conditions that banks seek to satisfy under production (credit) uncertainty (Pope and Just 1996). Thus, one may call into question the reliability of the results regarding banks' technological characteristics broadly reported in the literature, since the *ex-post* estimates of banking technology are likely to be biased and inconsistent (Pope and Chavas 1994). However, the extent to which these concerns are relevant for policy analysis is an empirical question.

In this paper, we shed light on this issue by highlighting the fundamental differences between the estimation of banks' production technology under uncertainty and that under the assumption of a deterministic production process. We first discuss why the underlying production process in banking—namely, the production of (income-generating) credit—is inherently uncertain.[2] Then, we review the *ex-post* modeling

[1] Some studies investigating banks' profitability account for liquidity risk using *ex-post* liquidity ratios (e.g., long-term loans to liquid liabilities, liquid assets to total assets, or liquid assets to deposits). See Shen et al. (2009) and references therein.

[2] Clearly, the fundamental principle of the production process in banking itself is quite certain, i.e., to borrow funds from one group of customers in the form of deposits and lend these funds to another group in the

approach commonly undertaken in the literature which either assumes uncertainty away or models it in a somewhat ad hoc fashion. We contrast this approach with its *ex-ante* counterpart which recognizes the uncertain nature of the production process in banking in the first stages of modeling. We show that the *ex-post* estimates of banking technology are likely to be biased and inconsistent and offer an alternative methodology to estimate banks' production technology based on the *ex-ante* (dual) cost function (Pope and Chavas 1994; Pope and Just 1996, 1998).[3] The latter acknowledges credit uncertainty explicitly by recognizing that bank managers minimize costs subject to given *expected* outputs and credit risk levels. We note that the estimation of the *ex-ante* cost function is advantageous over the primal approach to uncertain production processes (i.e., the expected utility maximization) because it is free of risk preference parameters and avoids the specification of the utility function. In order to make the model feasible to estimate, we estimate unobservable expected outputs and their associated credit risk levels from banks' supply functions (Moschini 2001) via nonparametric kernel methods.

We apply this *ex-ante* framework to data on U.S. commercial banks operating during the period from 2001 to 2010. The reported results on cost elasticities, scale economies, and productivity growth are contrasted with those obtained from the *ex-post* models of banking technology. We find that output elasticities of cost computed using the *ex-post* estimates of production technology tend to be biased upwards which in turn leads to downward biases in the returns to scale estimates. The results, however, do not differ qualitatively across the *ex-ante* and *ex-post* models if one controls for unobserved bank-specific effects. In the latter case, we find that virtually all U.S. commercial banks (regardless of the size) operate under increasing returns to scale (IRS), which is consistent with findings recently reported in the literature despite the differences in methodology (e.g.,Feng and Serletis 2010; Wheelock and Wilson 2012; Hughes and Mester 2013). Interestingly, if we leave bank-specific effects uncontrolled (as, for instance, the three above-cited studies do), the results change dramatically: the *ex-ante* models then indicate that 23–35 and 33–34 % of large banks exhibit decreasing and constant returns to scale (CRS), respectively.[4]

When analyzing the growth in total factor productivity (TFP) and its components, we find that, for medium and large banks, the TFP growth estimates from the *ex-ante* models tend to be higher than those from the *ex-post* models. The opposite is true for small banks. In fact, results from the *ex-ante* models indicate that the average annual TFP growth is negative among small banks. All models suggest that the bulk of the positive productivity growth in the industry comes from the scale economies component. According to the *ex-ante* models, the asset-weighted average annual TFP

Footnote 2 continued
form of loans. However, the amount of loans that will ultimately generate income for a bank is uncertain because not all issued loans are paid back duly. Since the fraction of the nonperforming loans is unknown to banks in advance, the latter makes the production of performing (earnings) loans uncertain.

[3] The terms "*ex-ante* cost function" and "*ex-post* cost function" were first coined by Pope and Chavas (1994).

[4] Restrepo-Tobón et al. (2013) similarly document that one is more likely to find the evidence of non-increasing returns to scale among commercial banks if unobserved effects are ignored.

growth due to IRS is around 2.1–2.2 % per annum across all banks. Despite that small banks exhibit higher economies of scale, we find that, on average, the TFP scale component is larger in magnitude for medium and large banks. Except for large banks, we find little evidence of economically significant technical progress. The level of expected credit risk, as measured by the volatility of earning assets, does not seem to impact productivity much either. For small and medium banks, we find no effect of expected risk levels on TFP growth; little negative effect is found for large banks.

The rest of the paper unfolds as follows. Section 2 discusses why the nature of banks' production process is inherently uncertain (stochastic) and describes the framework suitable to address this uncertainty in the estimation of banking technology. Section 3 provides a description of the data. The details of the estimation and the results are presented in Sect. 4. Section 5 concludes.

2 Stochastic banking technology

To avoid confusion, we note that, throughout the paper, we use the word "stochasticity" in its economic rather than econometric meaning, i.e., when saying "stochasticity" we mean "uncertainty" of the production process.

2.1 Credit uncertainty

We start by examining how the "production" is conceptually formalized in the case of banks. The framework broadly employed in the literature is the so-called "intermediation approach" of Sealey and Lindley (1977), according to which a bank's balance sheet is assumed to capture the essential structure of a bank's core business. Liabilities, together with physical capital and labor, are taken as inputs to the banks' production process, whereas assets (other than physical) are considered as outputs. Liabilities include core deposits and purchased funds; assets include loans and trading securities. Thus, this framework suggests the following mapping

$$\left\{\mathbf{x} \in \mathbb{R}_+^J;\ \mathbf{z} \in \mathbb{R}_+^K\right\} \quad \rightarrow \quad \left\{\mathbf{y} \in \mathbb{R}_+^M\right\}, \tag{1}$$

where $\mathbf{x}$ is a $J \times 1$ vector of variable inputs that include a bank's liabilities; $\mathbf{z}$ is a $K \times 1$ vector of (quasi-)fixed inputs[5] (if any); and $\mathbf{y}$ is an $M \times 1$ vector of outputs (assets and securities).

Numerous studies such as Berger and Mester (1997, 2003), Hughes and Mester (1998), Wheelock and Wilson (2001, 2012), Feng and Zhang (2012) among many others have used the above framework in their analysis of banking technologies. However, two points are worth mentioning here. First, most studies commonly specify total *issued* loans and securities as banks' outputs, whereas Sealey and Lindley (1977) argue at length that "only *earning* assets as outputs are consistent with rational profit maxi-

[5] Some studies of banks' production technologies also incorporate financial (equity) capital and income from off-balance-sheet activities as quasi-fixed netputs (e.g., Berger and Mester 1997, 2003). We address this issue in detail later in the paper.

mizing behavior" (p. 1260; emphasis added). Normally, not all issued loans are paid back duly. One may argue that it is more appropriate to consider only "performing" loans as earning assets and thus as banks' output. We recognize that Sealey and Lindley (1977) made the above argument largely to justify the inappropriateness of treating deposits as banks' outputs: loans, not deposits, earn banks income. In this paper, we, however, take Sealey and Lindley's (1977) argument a step further by tightening the definition of earning assets as those that actually earn income (i.e., performing loans) as opposed to merely have the potential to.[6] In fact, most existing studies do acknowledge the difference between the issued and performing loans (called nonperforming loans), as we will see below. Note that no such difference exists in Sealey and Lindley (1977) who explicitly assume that all assets are payable in full at maturity and that neither loans nor securities are subject to default (credit) risk (p. 1255). This brings us to the second point.

The original framework of Sealey and Lindley (1977) implies that the mapping in (1) is deterministic, since no assets are subject to credit risk. However, the latter assumption is rather strong and unrealistic. A desired alternative is to relax this assumption by allowing the bank's production process (1) to be stochastic (uncertain). By doing so, one would acknowledge that bank managers do not have perfect foresight and understand that some fraction of the issued loans may not be repaid. Note that this is consistent with the fact that banks actively engage in risk management which includes pre-screening of applicants, monitoring, collateralizing loans, and hedging. These risk mitigating activities are based on a bank's expected risk as opposed to realized risk. Formally, we define performing ($\mathbf{y}^+$) and nonperforming ($\mathbf{y}^-$) loans and securities such that $\mathbf{y} = \mathbf{y}^+ + \mathbf{y}^-$, where $\mathbf{y}$ is a vector of total issued loans. The bank's production process under credit uncertainty may then be represented as

$$\left\{\mathbf{x} \in \mathbb{R}_+^J;\ \mathbf{z} \in \mathbb{R}_+^K;\ \boldsymbol{\varepsilon} \in \mathbb{R}^M\right\} \rightarrow \left\{\mathbf{y}^+ \in \mathbb{R}_+^M\right\}, \tag{2}$$

where $\mathbf{y}^+$ is an $M \times 1$ vector of earning assets (i.e., performing loans and securities); and $\boldsymbol{\varepsilon}$ is an $M \times 1$ vector of corresponding mean-zero stochastic disturbances that represent the credit risk.

To clearly see the difference between the two formulations of the production process—deterministic (1) and stochastic (2)—we consider the banking technology represented by the primal production function. For the ease of discussion and notational simplicity, we use a simplified single-output representation of the bank's technology throughout this section. Then, the deterministic production function corresponding to the input-output mapping (1) takes the following form

$$y = f(\mathbf{x}, \mathbf{z}), \tag{3}$$

[6] This narrowed definition is consistent with a more realistic banks' objective, i.e., to maximize expected profits as opposed to maximize the potential for profits. It is also consistent with the proposition that bank managers (or banks themselves) maximize expected utility drawn from actual profits, not from the potential for profits.

where y is an output scalar, say, corresponding to total assets; $\mathbf{x}$ and $\mathbf{z}$ are as defined above; and $f(\cdot)$ is the production function.

Explicit modeling of credit risk leads to a stochastic production function, corresponding to (2), which can take the form

$$y^{+} = F(\mathbf{x}, \mathbf{z}, \varepsilon) \equiv f(\mathbf{x}, \mathbf{z}) \exp(\varepsilon), \tag{4}$$

where y^{+} is an output scalar corresponding to a total of *earning* assets; ε is an i.i.d. random disturbance with $\mathbb{E}[\varepsilon \,|\mathbf{x}, \mathbf{z}] = 0$ and $\mathbb{E}\left[\varepsilon^2 \,|\mathbf{x}, \mathbf{z}\right] = \sigma^2$. Equation (4) tells that output is not deterministically determined by a bank's inputs and may deviate from $f(\cdot)$ in the presence of a non-zero exogenous shock ε. The latter is more consistent with the reality that banks face than the assumption of no credit uncertainty implied by (3).

Modeling production as a stochastic (uncertain) process is not novel in economics (e.g., Feldstein 1971; Antle 1983); the approach is a common practice among agricultural economists [see Just and Pope (2002) and references therein]. Note that, while production function (4) allows for uncertainty in credit production by banks, it, however, assumes that the credit risk, as measured by the standard deviation of the output (a common measure of risk), is bank-invariant and cannot be influenced by banks. These two implications are too restrictive: (i) banks differ from one another in their riskiness and (ii) they actively engage in risk management thus (at least partly) influencing the magnitude of risk that they are exposed to. We can adopt the above factors into the production function as follows (Just and Pope 1978):

$$y^{+} = F(\mathbf{x}, \mathbf{z}, \varepsilon) \equiv f(\mathbf{x}, \mathbf{z}) + \varepsilon h(\mathbf{x}, \mathbf{z}), \tag{5}$$

where ε is an i.i.d. mean-zero, unit-variance random disturbance. However, according to equation (5), the volatility of output is no longer the same across banks and can now be affected by a bank's efforts as captured by the risk-management function $h(\cdot) \geq 0$. In other words, the standard deviation $\mathbb{S}\left[y^{+} \,|\mathbf{x}, \mathbf{z}\right] = \sigma h(\mathbf{x}, \mathbf{z}) \neq \sigma$.

One might think that, econometrically, the difference between the deterministic and stochastic formulations of the bank's production process [(1) and (2), respectively] is minuscule if one estimates the bank's production technology directly. The only difference between (3) and (5) is that the latter has the output defined as performing loans as opposed to total issued loans and that it has a heteroskedastic error. However, the difference will become more pronounced when one introduces the bank manager's behavior into the analysis (optimization and risk preferences), as we will see below. In what follows, we explicitly distinguish between deterministic and stochastic formulations of banking technology, as well as between the issued, performing and nonperforming loans (y, y^{+} and y^{-}, respectively).

2.2 The ex-ante cost function

When estimating banking technology in an attempt to obtain some metric of production technology such as economies of scale or scope, technical change, productivity or

efficiency, most studies in the literature use the dual approach (e.g., Hughes and Mester 1993, 1998, 2013; Hughes et al. 1996; Hughes et al. 2001; Wheelock and Wilson 2001, 2012; Feng and Serletis 2009, 2010; Restrepo-Tobón et al. 2013). By assuming that banks minimize costs, these studies are able to quantify banks' technology by estimating the dual cost function.[7] This is advantageous over the estimation of a primal specification of the production process mainly because it avoids the use of input quantities on the right-hand side of the regression equation which can lead to simultaneity (endogeneity) problems given that the input allocation is endogenous to a bank's decisions.

However, the majority of these studies appeal to a standard dual theory based on a deterministic production process like the one in (3), according to which the bank's dual cost function is defined as (in a single-output case)

$$C\left(y, \mathbf{w}, \mathbf{z}\right) = \min_{\mathbf{x}} \left\{ \mathbf{x}'\mathbf{w} \mid y \leq f\left(\mathbf{x}, \mathbf{z}\right);\ \mathbf{z} = \mathbf{z}_0 \right\}, \tag{6}$$

where C is the variable cost (cost of variable inputs); $\mathbf{w}$ is a vector of the competitive variable input prices; $\mathbf{z}$ is a vector of (quasi)-fixed inputs or "control variables" with the corresponding vector of observed (fixed) values $\mathbf{z}_0$; and the remaining arguments are as defined before.

Since the above method takes the risky nature of banking operations for granted, several attempts have been made to incorporate risk into the estimation of the banking technology. Hughes and Mester (1993, 1998) propose to condition the bank's cost function on financial (equity) capital and output quality (inverse of credit risk). By doing so, they allow (directly or indirectly) the price of uninsured deposits and the level of equity capital to be endogenous to a bank's decisions. The inclusion of the above variables into the cost function is motivated by a bank manager's utility maximization problem, according to which utility is a function not only of profits but also of output quality and equity capital. According to Hughes and Mester (1993, 1998), inclusion of output quality into a manager's utility function reflects the trade-off between profits and the credit risk associated with them. Equity capital may be a source of loanable funds, and thus can be used as a cushion against liquidity risk; it can also be a means of signaling the degree of a bank's credit riskiness to its depositors.

There may, however, be some concerns about implementing this method in practice. First, since the risk is not observed *ex ante*, researchers often resort to using its *ex-post* realizations. Output quality and risk are usually proxied by the *ex-post* ratio of nonperforming loans to total issued loans and by the average standard deviation of the bank's yearly net income during five prior years, respectively. Second, the underlying utility-maximization framework, based on which the dual cost function is ultimately defined in Hughes and Mester (1998), is still deterministic in its core. The latter amounts to an implicit assumption that risk and profits are known to bank managers *ex ante*, which might be rather strong.

[7] An alternative approach is to estimate the dual profit function under the premise of profit maximization. This is mostly popular amid the studies of inefficiency in the banking industry in the stochastic frontier framework (e.g., Berger and Humphrey 1997).

A more general treatment of the banks' risky technology is offered by Hughes et al. (1996); Hughes et al. (2000, 2001) who propose a model in which bank managers rank production plans (i.e., the set of input mix, output level and quality, and the level of equity capital) and profits according to their risk preferences and subjective conditional probability distribution of states of the world. The model is then estimated using Deaton and Muellbauer's (1980) almost ideal demand system which produces the "most-preferred" cost function. However, the concerns we express above are still likely to apply. While conditioning the bank managers' utility-maximization problem on their subjective probability distributions does free the model from stochasticity, it, however, comes at a cost. The model would then include expected values (and possibly higher moments) of all inherently stochastic arguments of the utility function which are unobserved at the time of decision-making. That is, managers would rank production plans based on the *expected* outputs, risk, and profits. Given that the latter variables are unobserved, one may instead use their *ex-post* realizations in the estimation.[8] Doing the latter would be equivalent to estimating banks' *ex-post* cost function which, however, may not reflect the actual optimality condition that bank managers seek to satisfy under production uncertainty (Pope and Just 1996).

An alternative approach to modeling banks' risky technology would be to recognize the uncertainty associated with the credit production by letting bank managers maximize expected utility subject to appropriate economic constraints. The principal drawback of this approach, however, is the need to either specify managers' utility function or, under certain conditions, its mean-variance representation in order to quantify banking technology (e.g., Chavas 2004). To avoid this, we instead suggest invoking the duality (under uncertainty) and recovering a bank's technology from its *ex-ante* cost function which is free of bank managers' risk preference parameters (Pope and Chavas 1994; Pope and Just 1996; Moschini 2001).

We rely on results by Pope and Chavas (1994) who show that an *ex-ante* cost function of the following form (in our notation)

$$C(\mathbf{G}, \mathbf{w}, \mathbf{z}) = \min_{\mathbf{x}} \left\{ \mathbf{x}'\mathbf{w} \mid \mathbf{G} \leq \mathbf{g}(\mathbf{x}, \mathbf{z});\ \mathbf{z} = \mathbf{z}_0 \right\} \tag{7}$$

is consistent with (the bank managers') expected utility maximization if and only if the revenue function takes the form $R(\mathbf{g}(\mathbf{x}, \mathbf{z}), \varepsilon, \cdot)$, where ε is a stochastic error in the production process $F(\mathbf{x}, \mathbf{z}, \varepsilon)$, and $\mathbf{g}(\cdot)$ is an $S \times 1$ vector of non-random constraints with the corresponding vector of levels $\mathbf{G}$. This implies that the cost minimization is interpreted as the first stage in a two-step decomposition of expected utility maximization (where the expected utility is set to be a function of uncertain profits). Intuitively, the above proposition says that the constraints $\mathbf{g}(\cdot)$ need to hold both expected revenue (i.e., expected output, given there is no price uncertainty) and the risk premium constant (Pope and Chavas 1994, p. 200).

In particular, under the assumption that the banking production technology takes the form in (5), where a bank's risk-management efforts are explicitly accounted for, the *ex-ante* cost function consistent with the bank managers' expected utility maximization is

[8] As, for instance, Hughes et al. (1996); Hughes et al. (2000, 2001) do.

$$C\left(G_1, G_2, \mathbf{w}, \mathbf{z}\right) = \min_{\mathbf{x}} \left\{ \mathbf{x}'\mathbf{w} \mid G_1 \leq f(\mathbf{x}, \mathbf{z});\ G_2 \leq h(\mathbf{x}, \mathbf{z});\ \mathbf{z} = \mathbf{z}_0 \right\}, \quad (8)$$

which is equivalent to

$$C\left(\overline{y}^+, \overline{\sigma}_{y^+}, \mathbf{w}, \mathbf{z}\right) = \min_{\mathbf{x}} \left\{ \mathbf{x}'\mathbf{w} \;\middle|\; \begin{array}{l} \overline{y}^+ \leq \mathbb{E}\left[F(\mathbf{x}, \mathbf{z}, \varepsilon)\right]; \\ \overline{\sigma}_{y^+} \leq \left(\mathbb{E}\left[F(\cdot) - \mathbb{E}[F(\cdot)]\right]^2\right)^{1/2}; \\ \mathbf{z} = \mathbf{z}_0 \end{array} \right\}. \quad (9)$$

The equivalence holds because $\mathbb{E}\left[F(\mathbf{x}, \mathbf{z}, \varepsilon)\right] = f(\mathbf{x}, \mathbf{z})$ and $\left(\mathbb{E}\left[F(\cdot) - \mathbb{E}[F(\cdot)]\right]^2\right)^{1/2} = \sigma h(\mathbf{x}, \mathbf{z})$, where the latter is proportional to $h(\mathbf{x}, \mathbf{z})$ [for details, see Pope and Chavas (1994)].

Equation (9) says that under credit uncertainty banks minimize cost holding expected output and expected standard deviation of output (that is, risk) constant. We note that the *ex-ante* cost minimization would be constrained by expected output only [regardless of the functional form of the production function $F(\mathbf{x}, \mathbf{z}, \varepsilon)$], if one is willing to assume a risk-neutral behavior by bank managers. This means that one should not justify the estimation of an *ex-post* cost function such as (6) or any modifications of it, which uses realized (*ex-post*) values of outputs and risk, by the assumption of risk-neutrality. We also emphasize that the cost C in (9) by no means excludes banks' expenses associated with actual nonperforming loans. The latter is consistent with bank managers optimally allocating inputs $\mathbf{x}$ *ex ante* based on their expected quantity of performing loans $\overline{y}^+$ (and the expected volatility in this quantity $\overline{\sigma}_{y^+}$), because one may not know in advance which loans that banks issue would eventually become nonperforming *ex post*.

To demonstrate the implication of estimating the *ex-post* as opposed to *ex-ante* cost function, we consider a simple example of a stochastic production function taking the form in (4). If we assume a single-input production (i.e., $\mathbf{x}$ is a scalar and $\mathbf{z} = \mathbf{0}$) and let the dual cost function take the translog form, the corresponding *ex-ante* cost function is

$$\ln C = \beta_0 + \beta_1 \ln \overline{y}^+ + \frac{1}{2}\beta_2 \left(\ln \overline{y}^+\right)^2 + \gamma_1 \ln w + \frac{1}{2}\gamma_2 (\ln w)^2 + \delta \ln \overline{y}^+ \ln w, \quad (10)$$

where $\overline{y}^+$ is the expected output.[9] However, as argued above, researchers traditionally estimate the *ex-post* cost function using the realized output. Thus, substituting $\overline{y}^+ = f(x) = y^+ \exp(-\varepsilon)$ from (4) into (10) yields

$$\ln C = \beta_0 + \beta_1 \ln y^+ + \frac{1}{2}\beta_2 \left(\ln y^+\right)^2 + \gamma_1 \ln w + \frac{1}{2}\gamma_2 (\ln w)^2 + \delta \ln y^+ \ln w + \xi, \quad (11)$$

where $\xi = \left\{-\beta_1\varepsilon + \frac{1}{2}\beta_2\varepsilon^2 - \beta_2\varepsilon \ln y^+ - \delta\varepsilon \ln w\right\}$ is the "error" that clearly is not mean-zero and is correlated with covariates through $\ln y^+$ and $\ln w$. Therefore, the estimates of banking technology produced by the *ex-post* cost function are likely to

[9] Here we use the narrow definition of earning assets as discussed above: performing loans (y^+), rather than total issued loans (y), are treated as the output.

be biased and inconsistent.[10] Moreover, Pope and Just (1996) show that the *ex-post* cost function may not necessarily have all standard properties of cost functions such as concavity in input prices; however, all properties apply to the *ex-ante* cost function.

3 Data

The data we use come from Call Reports publically available from the Federal Reserve Bank of Chicago. We include all FDIC insured commercial banks with reported data for 2001:I–2010:IV. We exclude internet banks, commercial banks conducting primarily credit card activities and banks chartered outside the continental U.S. We also omit observations for which negative values for assets, equity, outputs, off-balance-sheet income, and prices are reported. The resulting data sample is an unbalanced panel with 64,581 bank-year observations for 7,535 banks. All nominal stock variables are deflated to 2005 U.S. dollars using the Consumer Price Index (for all urban consumers).

We follow the abovementioned "intermediation approach" and define the following earning outputs for the *ex-ante* cost function: performing consumer loans (y_1^+), performing real estate loans (y_2^+), performing commercial and industrial loans (y_3^+), and earning securities (y_4^+).[11] These output categories are essentially the same as those in Berger and Mester (1997, 2003). 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 specify two quasi-fixed netputs: off-balance-sheet income (inc) and equity capital (k) (e.g.,Berger and Mester 1997, 2003; Feng and Serletis 2009). We thus concur with Hughes and Mester's (1993, 1998) argument that banks may use equity capital as a source of loanable funds, and thus as a cushion against losses. We compute the price of inputs ($\mathbf{w} = \{w_j\}_{j=1}^5$) by dividing total expenses on each input by the corresponding input quantity. Similarly, the price of outputs ($\mathbf{p} = \{p_m\}_{m=1}^4$) is computed by dividing total revenues from each output by the corresponding output quantity. We discuss the use for output prices in Sect. 4. Lastly, total variable cost (C) equals the sum of expenses on each of the five variable inputs.[12] Table 1 presents summary statistics of the data used in the analysis. For details on the construction of the variables, see the Appendix.

4 Estimation and results

Under the assumption of risk-aversion and a bank's production process that explicitly accounts for bank-varying credit uncertainty and risk-management efforts as repre-

[10] Here we abstract from other potential sources of biases across both the *ex-post* and *ex-ante* cost functions, such as biases due to the misspecification of the model, etc.

[11] These earning outputs are computed by subtracting the value of nonperforming loans and securities from the corresponding reported total values, i.e., $y_m^+ = y_m - y_m^- \quad \forall \quad m = 1, \ldots, 4$.

[12] As previously discussed, note that the total variable cost (C) includes expenses associated with total issued loans and securities (both those which turn out being performing and nonperforming) because the bank managers allocate inputs *ex ante*, i.e., before they know which loans would eventually become nonperforming.

Table 1 Summary statistics

Variable	Percentiles					
	Mean	5th	25th	Median	75th	95th
C	43,464.41	886.20	2,307.05	4,589.50	9,972.00	44,823.99
y_1	79,372.97	550.08	1,976.37	4,371.84	9,894.39	47,035.32
y_1^+	78,605.61	546.24	1,963.25	4,347.32	9,846.48	46,899.32
y_2	393,148.62	5,001.36	19,583.04	48,586.56	120,325.87	556,182.82
y_2^+	381,885.68	4,938.08	19,349.13	47,944.53	118,409.87	543,611.25
y_3	250,842.81	2,650.06	8,063.20	16,817.83	37,114.16	172,501.73
y_3^+	247,633.80	2,615.50	7,953.38	16,595.01	36,723.73	170,489.53
y_4	345,821.85	4,126.95	12,813.31	26,365.19	57,121.30	256,123.06
y_4^+	345,627.16	4,126.95	12,812.85	26,351.50	57,098.80	256,123.05
w_1	52.99	35.93	43.67	50.12	58.99	79.90
w_2	0.3963	0.1012	0.1660	0.2427	0.3950	1.0462
w_3	0.0344	0.0164	0.0257	0.0336	0.0429	0.0534
w_4	0.0104	0.0019	0.0047	0.0082	0.0138	0.0259
w_5	0.0261	0.0104	0.0181	0.0247	0.0332	0.0462
p_1	0.0927	0.0546	0.0756	0.0871	0.0997	0.1296
p_2	0.0708	0.0536	0.0633	0.0699	0.0778	0.0910
p_3	0.0856	0.0391	0.0626	0.0801	0.1041	0.1433
p_4	0.0403	0.0194	0.0329	0.0404	0.0470	0.0574
inc	17,613.76	19.00	92.00	284.00	969.00	7,784.00
k	116,088.34	2,557.5	6,101.5	11,611.75	24,281.00	107,893.75
$Assets$	1,197,667.50	23,324.75	57,255.75	112,795.52	245,347.16	1,115,980.10

The variables are defined as follows. C—total variable cost; y_1 and y_1^+—total issued and performing consumer loans, respectively; y_2 and y_2^+—total issued and performing real estate loans, respectively; y_3 and y_3^+—total issued and performing commercial and industrial loans, respectively; y_4 and y_4^+—total and earning securities, respectively; w_1—price of labor; w_2—price of physical capital; w_3—price of purchased funds; w_4—price of interest-bearing transaction accounts; w_5—price of non-transaction accounts; p_1, p_2, p_3 and p_4—prices of y_1^+, y_2^+, y_3^+ and y_4^+, respectively; inc—off-balance-sheet income; k—equity capital; $Assets$—total assets. All variables but input and output prices are in thousands of real 2005 US dollars. All input and output prices but w_1 are interest rates and thus are unit-free

sented by (5), the estimation of the *ex-ante* cost function (9) requires the knowledge of bank managers' *expectations* of the levels of earning outputs ($\overline{\mathbf{y}}^+$) and associated credit risk levels as captured by the standard deviations ($\overline{\boldsymbol{\sigma}}_{\mathbf{y}^+}$). These expected outputs and volatility are not observed but can be estimated. For instance, Pope and Just (1996, 1998) suggest letting the *ex-ante* cost function take a self-dual functional form so that one can recover the underlying production function in its closed form. One can then use the recovered production function to estimate the expected value of outputs (i.e., conditional mean) by regressing their *ex-post* realizations on inputs.[13] Flexible

[13] Note that one does not need to regress outputs on inputs in order to obtain expected output levels *per se*: they can rather be recovered indirectly inside the numerical optimization algorithm (see Pope and Just 1996 for details). This approach, however, is still subject to Moschini's (2001) criticism.

functional forms such as the translog or Fourier clearly do not belong to the family of such self-dual cost functions. Further, this approach is subject to criticism, since it introduces endogeneity into the estimation because inputs are endogenous to banks' cost-minimizing decisions (for details, see Moschini 2001).

We do not wish to restrict our analysis to an inflexible self-dual specification of the *ex-ante* cost function, and therefore, follow an alternative approach. We opt to recover bank managers' expectations of outputs and risk levels (as measured by volatility) via kernel methods as suggested by Pope and Chavas (1994). This allows us to let the cost function take any desired form, particularly the translog which is widely used in the banking literature. Thus, the estimation of the *ex-ante* cost function consists of two stages: (i) the estimation of the unobserved expected outputs and associated credit risk levels and (ii) the estimation of banks' production technology via cost function.

In the first stage, we estimate the expectations of banks' outputs and their associated credit risk levels via nonparametric kernel methods. In order to avoid the introduction of endogeneity by estimating the production function as described above, here we follow Moschini's (2001) advice and instead estimate the corresponding (nonparametric) supply functions that are functions of exogenous output and input prices, and quasi-fixed inputs: $y_m^+ = s_m(\mathbf{p}, \mathbf{w}, k) \quad \forall \quad m = 1, \ldots, 4$, where the subscript m designates one of the four outputs. We use the local-constant least squares estimator[14] to estimate the bank-year-specific expected outputs ($\overline{y}_m^+$), i.e.,

$$\overline{y}_{m,it}^+(d) = \left[\sum_{i=1}^{N}\sum_{t=1}^{T} K\left(\frac{\mathbf{D}_{it} - d}{\mathbf{h}_m}\right)\right]^{-1}\left[\sum_{i=1}^{N}\sum_{t=1}^{T} y_{m,it}^+ K\left(\frac{\mathbf{D}_{it} - d}{\mathbf{h}_m}\right)\right], \tag{12}$$

where $\mathbf{D}_{it} \equiv (\mathbf{p}_{it}, \mathbf{w}_{it}, k_{it})$ is a vector of arguments of the supply function[15]; $K(\cdot)$ is a product kernel function (Racine and Li 2004);[16] and $\mathbf{h}_m$ is a vector of optimal bandwidths which we select via the data-driven least squares cross-validation (LSCV) method (Li and Racine 2004).[17] We divide the output and input prices by the price of one of the outputs in order to impose the homogeneity of degree zero onto the supply function.

[14] We opt for the local-constant estimator as opposed to the local-polynomial estimator (which has the same asymptotic variance but smaller bias) because the selection of optimal bandwidths for the former is less computationally demanding than it is for the alternative. The latter is a non-negligible issue given a large sample size of the dataset we use, as well as the number of estimations we need to perform.

[15] In order to control for year and fixed effects, we also include the time trend, as well as an unordered bank-index variable. This is similar in nature to the least squares dummy variable approach in parametric panel data models with fixed effects.

[16] We use a second-order Gaussian kernel for continuous covariates and Racine and Li's (2004) kernels for ordered and unordered covariates (i.e., the time trend and the bank index, respectively).

[17] Given that the cross-validation (CV) function is often not smooth in practice, we use multiple starting values for bandwidths when optimizing the function in order to ensure a successful convergence. Also, although we use constant bandwidths in our analysis, we acknowledge that one may instead prefer the use of adaptive bandwidths which adjust to the local sparseness of the data (if there is any). The selection of the adaptive bandwidths would, however, be more computationally demanding, especially given a relatively large number of dimensions in which the CV function needs to be optimized.

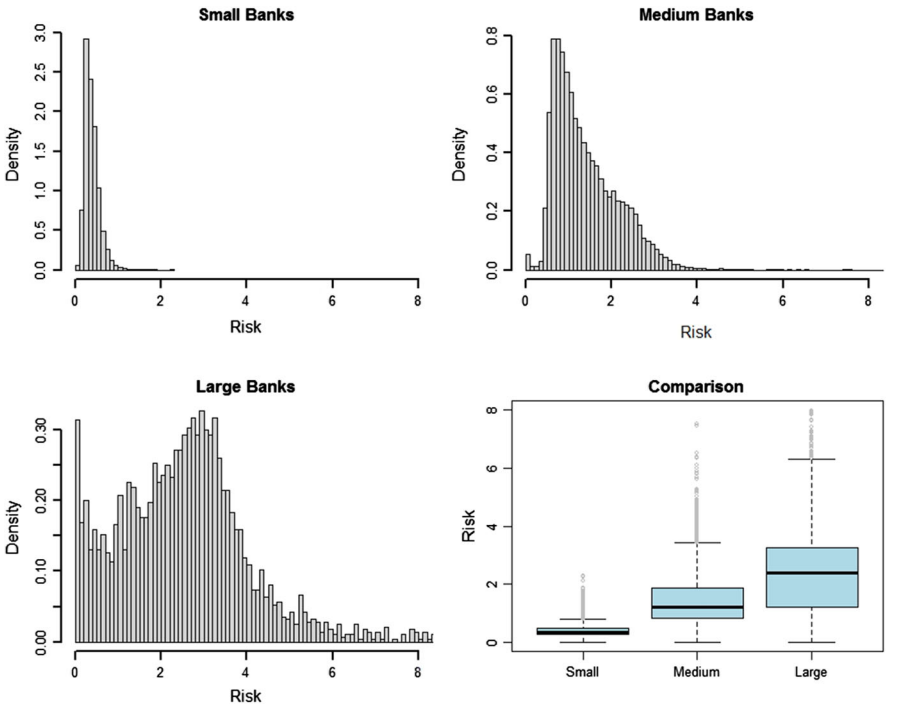

Fig. 1 Histograms of output-weighted credit risk estimates

We obtain the estimates of expected credit risk level ($\overline{\sigma}_{y_m^+}$ corresponding to each of the four outputs, as standardly measured by their volatility, by nonparametrically regressing the squares of residuals from (12) on $\mathbf{D}_{it}$. The latter produces the bank-year-specific estimates of conditional variances of the four outputs (under the mean-zero assumption for the error terms).[18] Similarly, here we use the LSCV to select optimal bandwidths. Note that the use of the local-constant estimator automatically ensures that no negative fitted values of the variance are produced. Figure 1 presents histograms of these expected credit risk estimates tabulated by the bank-size category. We classify a bank as "small" if its total assets are below $100 million, "medium" if total assets are between $100 million and $1 billion, and "large" if total assets exceed $1 billion. Here, we plot an output-weighted expected risk (scaled down by its standard deviation) in order to avoid plotting four different risk measures associated with each of the four outputs that we consider. As expected, the distributions are positively skewed with mean (median) risk increasing with the size of the bank. The fit in both the estimation of $\overline{\mathbf{y}}^+$ and $\overline{\boldsymbol{\sigma}}_{\mathbf{y}^+}$, measured by the square of the correlation coefficient between the actual and predicted values, is as high as 0.99.[19]

[18] McAllister and McManus (1993) use a similar nonparametric procedure to estimate the expected rate of return and associated expected risk for U.S. banks.

[19] To conserve space, we do not report the detailed results from the first stage (they are available upon request) and directly proceed to the discussion of the main results from the second stage.

In the second stage, we estimate the multi-output generalization of the *ex-ante* cost function under risk-aversion (9) which takes the following form (under the translog specification)[20]

$$
\begin{aligned}
\ln C_{it} = {} & \alpha_0 + \sum_{m=1}^{4} \alpha_m \ln \overline{y}^{+}_{m,it} + \frac{1}{2} \sum_{m=1}^{4} \sum_{h=1}^{4} \alpha_{mh} \ln \overline{y}^{+}_{m,it} \ln \overline{y}^{+}_{h,it} \\
& + \sum_{m=1}^{4} \beta_m \ln \overline{\sigma}_{y^{+}_{m},it} + \frac{1}{2} \sum_{m=1}^{4} \sum_{h=1}^{4} \beta_{mh} \ln \overline{\sigma}_{y^{+}_{m},it} \ln \overline{\sigma}_{y^{+}_{h},it} \\
& + \sum_{m=1}^{4} \sum_{h=1}^{4} \gamma_{mh} \ln \overline{y}^{+}_{m,it} \ln \overline{\sigma}_{y^{+}_{h},it} \\
& + \sum_{j=1}^{5} \delta_j \ln w_{j,it} + \frac{1}{2} \sum_{j=1}^{5} \sum_{s=1}^{5} \delta_{js} \ln w_{j,it} \ln w_{s,it} \\
& + \sum_{m=1}^{4} \sum_{j=1}^{5} \eta_{mj} \ln \overline{y}^{+}_{m,it} \ln w_{j,it} + \sum_{m=1}^{4} \sum_{j=1}^{5} \mu_{mj} \ln \overline{\sigma}_{y^{+}_{m},it} \ln w_{j,it} \\
& + \sum_{l=1}^{2} \theta_l \ln z_{l,it} + \frac{1}{2} \sum_{l=1}^{2} \sum_{r=1}^{2} \theta_{lr} \ln z_{l,it} \ln z_{r,it} + \omega_1 t + \frac{1}{2} \omega_{11} t^2 \\
& + \sum_{m=1}^{4} \sum_{l=1}^{2} \varphi_{ml} \ln \overline{y}^{+}_{m,it} \ln z_{l,it} + \sum_{m=1}^{4} \sum_{l=1}^{2} \phi_{ml} \ln \overline{\sigma}_{y^{+}_{m},it} \ln z_{l,it} \\
& + \sum_{j=1}^{5} \sum_{l=1}^{2} \psi_{jl} \ln w_{j,it} \ln z_{l,it} \\
& + \sum_{m=1}^{4} \pi_{m1} \ln \overline{y}^{+}_{m,it}\, t + \sum_{m=1}^{4} \pi_{m2} \ln \overline{\sigma}_{y^{+}_{m},it}\, t + \sum_{j=1}^{5} \pi_{j3} \ln w_{j,it}\, t \\
& + \sum_{l=1}^{2} \pi_{l4} \ln z_{l,it}\, t + \lambda_i, \qquad (13)
\end{aligned}
$$

where $\mathbf{z} = (inc, k)'$; t is the time trend; λ_i is the bank-specific (unobserved) fixed effect; and the remaining variables are as defined before. To account for unobserved heterogeneity, we include bank-specific fixed effects (λ_i) in the above cost function. The latter is broadly overlooked in the literature which might lead to biased and

[20] Note that, like in the first stage, the *ex-ante* cost function could have been alternatively estimated via nonparametric kernel methods. In this paper, we however, opt for (admittedly more restrictive) translog specification. We acknowledge that several papers have documented that the translog form may sometimes be a poor approximation of banks' cost function (e.g., Wheelock and Wilson 2001, 2012). We nevertheless opt for this parametric specification in order to facilitate the comparison of our findings with the results in the existing banking literature that overwhelmingly favors the translog specification.

misleading results (see the discussion in Restrepo-Tobón et al. 2013). We also include the time trend to capture time effects/technical change.

In order to analyze how the use of (i) different approaches to accommodate credit uncertainty in the analysis of banks' cost technologies and (ii) different definitions of banks' earning outputs (total issued loans and securities versus performing loans and securities) affects the conclusions that researchers draw about the banking technology, we estimate a number of auxiliary models in addition to the *ex-ante* cost function under risk-aversion in (13). For the ease of discussion, below we define all second-stage models we estimate.

Model I The model of banking technology under credit uncertainty and risk-aversion in the form of the *ex-ante* cost function (13).

Model II The model of banking technology under credit uncertainty and risk-neutrality in the form of the *ex-ante* cost function. Here, we estimate (13) with the expected risk measures omitted from the equation (see the discussion in Sect. 2.2).

Model III The model of banking technology estimated via the *ex-post* cost function. Here, we estimate (13) with realized values of *performing* loans and securities ($\mathbf{y}^+$) used as earning outputs (as opposed to expected values $\bar{\mathbf{y}}^+$) and the expected risk measures omitted from the equation.

Model IV The model of banking technology estimated via the *ex-post* cost function with total *issued* loans and securities ($\mathbf{y}$) used as earning outputs. No risk measures are included. This is the most commonly estimated model in the banking literature.

Model V The model of banking technology estimated via the *ex-post* cost function with total *issued* loans and securities ($\mathbf{y}$) used as earning outputs. Here, we add an *ex-post* measure of credit risk as proxied by the ratio of total non-performing loans to issued loans, widely used in the literature to "control" for risk in an ad hoc manner.

The fixed effect adjustment is done via the within transformation. For all model specifications, we estimate the SUR system consisting of the cost function and the corresponding input cost share equations, onto which we impose the symmetry, linear homogeneity (in input prices), and cross-equation restrictions.

4.1 Elasticities

We first examine the differences between the models in terms of implied elasticities of banks' costs. Table 2 reports average elasticity estimates for outputs, input prices, quasi-fixed netputs, and the time trend from all five models, based on which we can make several observations.

As expected, estimates of input price elasticities do not differ across the models, since they are constrained by the same input cost share equations. However, when examining output elasticities, we find that, while there are little differences within the groups of *ex-ante* (I and II) and *ex-post* (III–V) models, there are dramatic changes across the two groups. In the case of all four outputs, the *ex-ante* Models I and II report average elasticities that are significantly smaller than those from the *ex-post*

Table 2 Mean elasticity estimates

Variable	I	II	III	IV	V
output 1	0.023	0.022	0.064	0.064	0.064
output 2	0.397	0.386	0.407	0.408	0.407
output 3	0.097	0.098	0.142	0.141	0.141
output 4	0.060	0.061	0.170	0.169	0.169
w_1	0.413	0.413	0.412	0.412	0.412
w_2	0.102	0.102	0.101	0.101	0.101
w_3	0.167	0.167	0.168	0.168	0.168
w_4	0.026	0.026	0.026	0.026	0.026
w_5	0.292	0.292	0.294	0.294	0.294
k	0.071	0.081	0.075	0.079	0.079
inc	0.076	0.077	0.033	0.033	0.033
t	0.000	0.000	0.000	−0.001	−0.002

While the output categories are the same across five models, the values used in the estimation are different. Models I and II use estimates of the expected performing loans and securities; Model III uses the realized values of performing loans and securities; and Models IV and V use the realized values of issued loans and securities (performing + nonperforming). See the text for details

Models III–V. For instance, the estimated elasticity of cost with respect to consumer loans (*output* 1) from Models I and II is, on average, 2.9 times smaller than its counterpart obtained from Models III to V. The difference is of similar magnitude in the case of securities (*output* 4). Notably, all five models show that banks' cost is the most sensitive to changes in the level of real estate loans (*output* 2). However, the two groups differ in the second most "cost-influential" output: commercial loans (*output* 3) according to Models I and II versus securities (*output* 4) as predicted by Models III–V.

For equity capital (k), the differences in mean elasticities are minuscule. We consistently find it to be positive across the models, which leads us to conclude that banks do not rely on financial capital as a source of loanable funds, but rather consider it as an "output." This is in line with Hughes and Mester's (1998) argument that banks might use equity capital as a means of signaling their overall riskiness to customers.[21]

4.2 Scale economies

Table 3 presents the summary statistics of the point estimates of returns to scale based on all five models over the entire sample period.[22] We break down the results by the asset size of banks. Overall, we find the returns to scale estimates from the *ex-ante* Models I and II to be larger than those from Models III–V. This result was expected given our findings that the output elasticities, based on the results from *ex-ante* models, are consistently smaller than those from *ex-post* models. Recall that returns to scale estimates are the inverse of the sum of cost elasticities with respect to outputs. These

[21] This result may also stem from the fact that financial regulations require equity (financial) capital to expand in proportion to loans.

[22] When computing these summary statistics, we omit the first and the last percentiles of the distribution of the returns to scale estimates, in order to minimize the influence of outliers. However, the omitted estimates correspond to the same observations across all five models, in order to keep the results comparable. We, therefore, can still cross-reference results from different models at the bank level.

Table 3 Summary of returns to scale estimates

	Point estimates of RS							Categories of RS		
	Mean	SD	Min	1st qu.	Median	3rd qu.	Max	DRS	CRS	IRS
	Small banks									
I	2.004	0.278	1.292	1.793	1.979	2.188	2.911	0	0	28,700
II	2.044	0.266	1.366	1.839	2.024	2.223	2.758	0	0	28,700
III	1.333	0.091	0.973	1.271	1.320	1.380	2.324	0	4	28,696
IV	1.337	0.091	0.967	1.276	1.325	1.385	2.311	1	3	28,696
V	1.332	0.091	0.944	1.269	1.318	1.380	2.094	2	4	28,694
	Medium banks									
I	1.633	0.208	1.072	1.482	1.610	1.765	2.908	0	0	31,656
II	1.656	0.199	1.152	1.512	1.636	1.787	2.656	0	0	31,656
III	1.246	0.072	0.901	1.198	1.238	1.285	2.714	4	5	31,647
IV	1.250	0.072	0.903	1.202	1.242	1.289	2.677	4	5	31,647
V	1.254	0.073	0.901	1.206	1.245	1.292	2.252	4	7	31,645
	Large banks									
I	1.313	0.150	1.084	1.203	1.284	1.394	2.613	0	0	2,935
II	1.332	0.147	1.134	1.227	1.305	1.410	2.617	0	0	2,935
III	1.180	0.100	1.021	1.127	1.166	1.210	2.934	0	0	2,935
IV	1.182	0.098	1.017	1.129	1.168	1.213	2.904	0	0	2,935
V	1.197	0.084	1.016	1.146	1.185	1.229	2.271	0	0	2,935
	All banks									
I	1.787	0.318	1.072	1.554	1.750	1.988	2.911	0	0	63,291
II	1.817	0.316	1.134	1.584	1.782	2.022	2.758	0	0	63,291
III	1.283	0.095	0.901	1.218	1.271	1.333	2.934	4	9	63,278
IV	1.287	0.096	0.903	1.222	1.275	1.337	2.904	5	8	63,278
V	1.287	0.093	0.901	1.224	1.274	1.335	2.271	6	11	63,274

downward biases in returns to scale estimates produced by the *ex-post* Models III–V are more apparent in Fig. 2 which plots kernel densities of these estimates.

The right pane of Table 3 also reports the groupings of banks by the returns to scale categories: decreasing returns to scale (DRS), CRS, and IRS. We classify a bank as exhibiting DRS/CRS/IRS if the point estimate of its returns to scale is found to be statistically less than/equal to/greater than one at the 95 % significance level.[23] The empirical evidence suggests that there is little (if any at all) qualitative difference in scale economies across the models: virtually all banks are found to exhibit IRS regardless whether the *ex-ante* or *ex-post* cost function is being estimated. These findings of IRS are consistent with those recently reported in the literature despite the differences in methodology (e.g.,Feng and Serletis 2010; Hughes and Mester 2013; Wheelock and Wilson 2012). However, examining the Spearman's rank correlation coefficients of the scale economies estimates across the models (see Table 4) suggests

[23] Standard errors are constructed using the delta method.

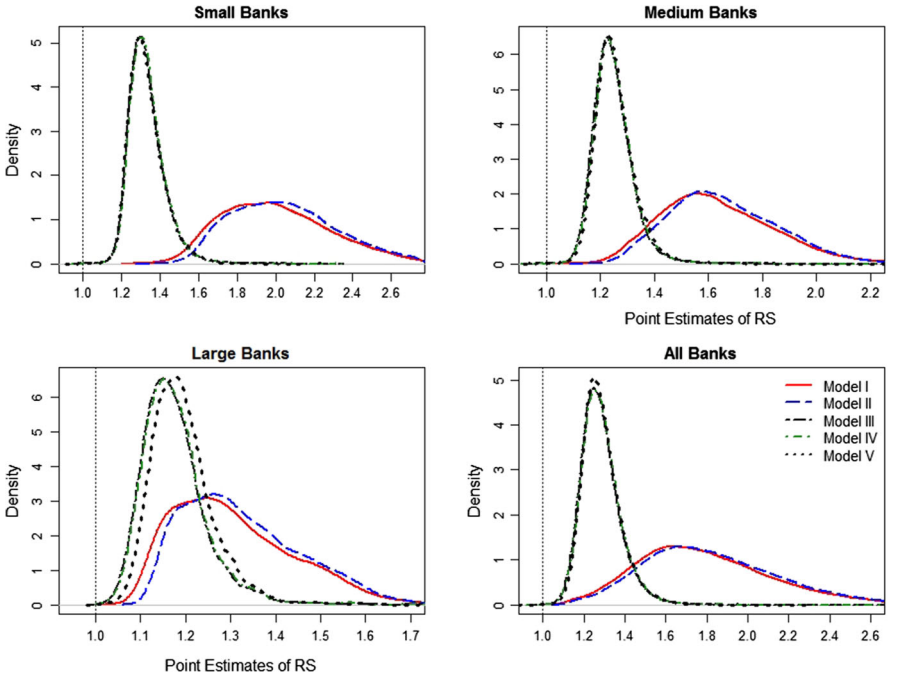

Fig. 2 Kernel densities of returns to scale estimates

striking differences in rankings of the banks. The rank correlation coefficient between the *ex-ante* models and *ex-post* models is around 0.3 for small banks, 0.45 for medium size banks, and 0.6 for large banks.[24]

Notably, the results on economies of scale change dramatically if we do not account for unobserved bank-specific heterogeneity when estimating the models, as commonly done in the literature (e.g., Hughes et al. 2001; Wheelock and Wilson 2012). We find that biases in returns to scale estimates from the *ex-post* Models III–V are no longer uniformly negative across all bank-size categories. In particular, these models tend to *over*-estimate the scale economies for large banks. The *ex-post* Models III–V estimate that 88–89 % of large U.S. banks exhibit IRS, whereas the results from the *ex-ante* Models I and II indicate that 23–35 and 33–34 % of large banks exhibit DRS and CRS, respectively.[25]

4.3 Productivity

The estimation of the cost function allows the econometric decomposition of the total factor productivity (TFP) growth (defined as $T\dot{F}P = \dot{y} - \sum_{j=1}^{J} s_j \dot{x}_j$ in a single-

[24] We note that one ought to be careful here when comparing rankings using the Spearman's rank correlation coefficient because the latter does not account for the estimation error associated with the estimation of scale economies.

[25] To conserve space, we do not report detailed results from these models; they are available upon request.

Table 4 Rank correlation coefficients of returns to scale estimates

Model	I	II	III	IV	V	I	II	III	IV	V
	Small banks					*Large banks*				
I	1.00					1.00				
II	0.99	1.00				0.99	1.00			
III	0.31	0.34	1.00			0.57	0.58	1.00		
IV	0.33	0.36	0.99	1.00		0.60	0.61	0.99	1.00	
V	0.30	0.34	0.91	0.91	1.00	0.60	0.61	0.96	0.97	1.00
	Medium banks					*All banks*				
I	1.00					1.00				
II	0.99	1.00				0.99	1.00			
III	0.42	0.44	1.00			0.59	0.61	1.00		
IV	0.45	0.47	0.99	1.00		0.61	0.63	0.99	1.00	
V	0.44	0.45	0.94	0.95	1.00	0.57	0.59	0.95	0.95	1.00

output case[26]) into several components (e.g.,Denny et al. 1981). However, the latter procedure is designed for production processes under certainty, which is clearly not the case in our study. For instance, Solow (1957) derives this Divisia index assuming that the production process is deterministic. In order to be able to follow his derivation in the presence of uncertainty, one first needs to take the expected values of both sides of the production function. It is easy to show that in this case, the Divisia index of the TFP growth (with a single output) will be defined as $T\dot{F}P = \dot{\bar{y}} - \sum_{j=1}^{J} s_j \dot{x}_j$, where the growth in *expected* output ($\bar{y}$) rather than actual output is used. Naturally, one can claim that asymptotically the two measures will be the same, since the average of errors approaches zero as $N \to \infty$. However, the latter will not hold true if the variance of the error conditional on inputs is not constant *over time*, as we have in our case where banks are explicitly allowed to manage risk through the time-dependent $h(\cdot)$ function in (5). Further, as we have discussed in Sect. 2, the choice of outputs is ambiguous in the case of banks. Researchers have customarily used total issued loans and securities as banks' earning outputs, whereas we employ a more narrow definition in this paper by considering performing loans and securities only as banks' outputs. We, therefore, compute three different Divisia indices of the TFP growth.[27]

Divisia 1 The Divisia index of the TFP growth computed using expected earning outputs ($\bar{\mathbf{y}}^{+}$). This index is comparable to the TFP estimates from the *ex-ante* Models I and II.

Divisia 2 The Divisia index of the TFP growth computed using realized values of earning outputs ($\mathbf{y}^{+}$). This index is comparable to the results from the *ex-post* Model III.

[26] Here, the "dot" designates the growth rate and s_j is the cost share of the jth input.

[27] Since we have four outputs, we follow the literature and use the revenue-shared weighted output growth when computing the TFP growth, i.e., $T\dot{F}P = \sum_{m=1}^{4} r_m \dot{y}_m - \sum_{j=1}^{5} s_j \dot{x}_j$, where r_m is the revenue share of the mth output.

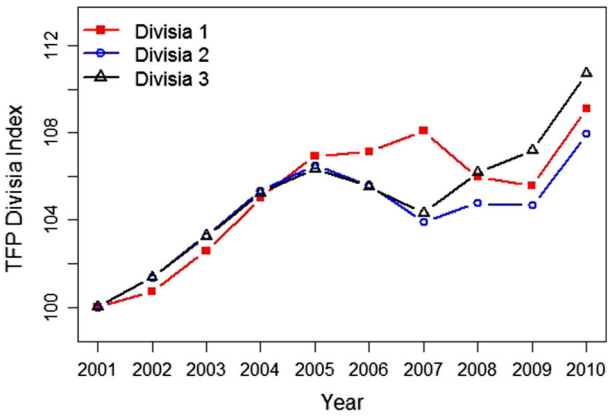

Fig. 3 TFP Divisia indices

Divisia 3 The Divisia index of the TFP growth computed using realized values of outputs that include nonperforming loans and securities (**y**). This index is comparable to the results from the *ex-post* Models IV and V.

Figure 3 plots these indices normalized to a hundred in 2001.[28] The three indices are virtually indistinguishable up until 2005, when the divergence between the Divisia index 1 and the remaining two indices starts. The Divisia 1 indicates that banks had higher expectations of productivity growth up until the onset of the financial crisis in 2007. In 2008, however, the expectation-based productivity growth plunges down. Notably, there has been a growing gap between index 2 and 3 since 2007. We attribute this to the inability of the Divisia index 3 to account for a large share of nonperforming loans that banks have been handling after the crisis. Overall, we estimate the productivity in the industry to have grown by 9 % over the 2001–2010 period.

As noted before, we can decompose the TFP growth by estimating its components econometrically from the dual cost function. By adding and subtracting the Divisia index 1 from the *ex-ante* cost function (13) of Model I and then totally differentiating the latter with respect to time, it is easy to show that the TFP growth can be decomposed into

$$\dot{TFP} = \sum_m \left(r_m - \frac{\partial \ln C\left(\overline{\mathbf{y}}^+, \overline{\boldsymbol{\sigma}}_{\mathbf{y}^+}, \mathbf{w}, \mathbf{z}\right)}{\partial \ln \overline{y}_m^+} \right) \dot{\overline{y}}_m^+ + \sum_j \left(s_j - \frac{\partial \ln C(\cdot)}{\partial \ln w_j} \right) \dot{w}_j - \sum_l \frac{\partial \ln C(\cdot)}{\partial \ln z_l} \dot{z}_l - \sum_m \frac{\partial \ln C(\cdot)}{\partial \ln \overline{\sigma}_m^+} \dot{\overline{\sigma}}_m^+ - \frac{\partial \ln C(\cdot)}{\partial \ln t}, \tag{14}$$

where r_m is the revenue share of the mth output; s_j is the cost share of the jth input; and the remaining variables are as defined before. The "dot" designates the growth rate.

[28] Since the Divisia index is bank specific, in order to construct the plot, we use the asset-weighted average annual TFP growth rates.

The components in (14) are defined as follows [in order of appearance]: (i) "scale component" which can be shown to consist of two subcomponents which depend on returns to scale and the mark-up (departure of output prices from their respective marginal costs); (ii) "allocative component" which captures the effects of non-optimal input allocation; (iii) "exogenous component" which captures effects of quasi-fixed netputs (here, off-balance-sheet income and equity capital); (iv) "risk component" which accounts for the effect of the risk levels; and (v) "technical change" that captures temporal shifts in the estimated cost function.[29] Clearly, the definition of the outputs will be changing with the model: $\overline{y}^{+}$ versus y^{+} versus y. Further, we will have the risk component only when estimating Models I and V (recall they are the only ones that control for risk).[30]

We report the TFP growth components across the models in Table 5. For medium and large banks, the TFP growth estimates from the *ex-ante* Models I and II are higher than those from the *ex-post* Models III–V. The opposite is true for small banks. In fact, we find that the average annual TFP growth is negative for small banks, based on Models I and II.

We consistently find the exogenous component to negatively contribute (about $-1\,\%$) to the TFP growth across all five models and all bank-size categories. There is also some evidence of negative effects of input misallocation on TFP in the case of large banks. All models suggest that the bulk of the positive productivity growth comes from the scale economies component. The *ex-ante* Models I and II estimate the asset-weighted average annual TFP growth due to IRS to be around 2.1–2.2 % per annum across all banks. The estimate is about 1.5 times smaller based on the *ex-post* Models III–V (around 1.5–1.6 % per annum for all banks). Notably, despite that small banks exhibit the largest economies of scale in our sample (see Table 3), the results from Models I and II indicate that the scale component is larger in magnitude for medium and large banks instead.

We find little evidence of economically significant technical progress, except for large banks according to the *ex-ante* Models I and II (about 1.6–1.7 % per annum, on average). Risk level does not seem to impact productivity much either. Based on Model I, we find no effect of expected risk levels on small and medium banks' productivity growth; little negative effect is found for large banks (average annual rate of $-0.8\,\%$).

We use the estimated TFP growth rates to construct the TFP indices (normalized to a hundred in 2001) which we plot in Fig. 4 with the corresponding Divisia indices. The TFP indices from all five models are compared to Divisia 1 and 2, which we believe are more reliable measures of productivity change, since they are based on *performing* loans and securities. We find that the Divisia indices generally lay in between the indices based on the estimates from the group of *ex-ante* and the group of *ex-post* models. Note that we do not seek to compare the indices obtained based on econometric models with the data-based (nonparametric) Divisia indices. Had our goal been to (econometrically) estimate the TFP index that is equal to the Divisia

[29] For more on the decomposition of the TFP growth, see Kumbhakar and Lovell (2000).

[30] In the case of the ex-post Model V, it is easy to show that the risk component of the TFP growth is defined as the negative of $\frac{\partial \ln C(\cdot)}{\partial \ln npl}\dot{npl}$, where $\dot{npl}$ is the growth rate of the ratio of total nonperforming loans to issued loans (an *ex-post* measure of credit risk).

Table 5 Weighted average annual growth in TFP and its components

Model	TC	Scale	Allocative	Exogenous	Risk	Total
	Small banks					
I	0.0000	0.0080	0.0021	−0.0112	0.0000	−0.0012
II	0.0001	0.0084	0.0021	−0.0117	–	−0.0012
III	0.0005	0.0140	0.0021	−0.0090	–	0.0076
IV	0.0017	0.0144	0.0021	−0.0092	–	0.0091
V	0.0022	0.0149	0.0020	−0.0093	−0.0021	0.0078
	Medium banks					
I	−0.0024	0.0238	0.0009	−0.0108	0.0000	0.0115
II	−0.0022	0.0247	0.0009	−0.0118	–	0.0115
III	−0.0022	0.0149	0.0010	−0.0116	–	0.0021
IV	0.0000	0.0156	0.0010	−0.0118	–	0.0048
V	0.0012	0.0162	0.0010	−0.0122	−0.0020	0.0042
	Large banks					
I	0.0166	0.0218	−0.0040	−0.0134	−0.0079	0.0131
II	0.0153	0.0196	−0.0041	−0.0143	–	0.0165
III	0.0058	0.0143	−0.0021	−0.0111	–	0.0070
IV	0.0075	0.0141	−0.0021	−0.0110	–	0.0086
V	0.0084	0.0153	−0.0021	−0.0120	0.0002	0.0098
	All banks					
I	0.0071	0.0216	−0.0014	−0.0121	−0.0039	0.0113
II	0.0066	0.0209	−0.0015	−0.0130	–	0.0130
III	0.0020	0.0145	−0.0004	−0.0112	–	0.0049
IV	0.0039	0.0148	−0.0004	−0.0112	–	0.0070
V	0.0048	0.0157	−0.0005	−0.0119	−0.0009	0.0072

The estimates are obtained by averaging the bank-year specific annual TFP growth rates over the entire sample period using the total assets as weights

index, we could have imposed this equality in the estimation as, for instance, done by Kumbhakar and Lozano-Vivas (2005).

5 Conclusion

Risk is crucial to banks' production. Banks actively engage in risk assessment, risk monitoring, and other risk-management activities. Therefore, researchers should explicitly incorporate banks' risk-taking behavior when estimating their production technology. The banking literature has largely focused on the estimation of the *ex-post* realization of banking technology with credit risk being either completely overlooked or controlled for in a somewhat ad hoc manner. Most studies use *ex-post* realizations of risk-related variables and uncertain outputs. These methods, however, may not reflect the optimality conditions that bank managers seek to satisfy *ex ante* under credit uncertainty. One thus may call into question the reliability of the results from such studies.

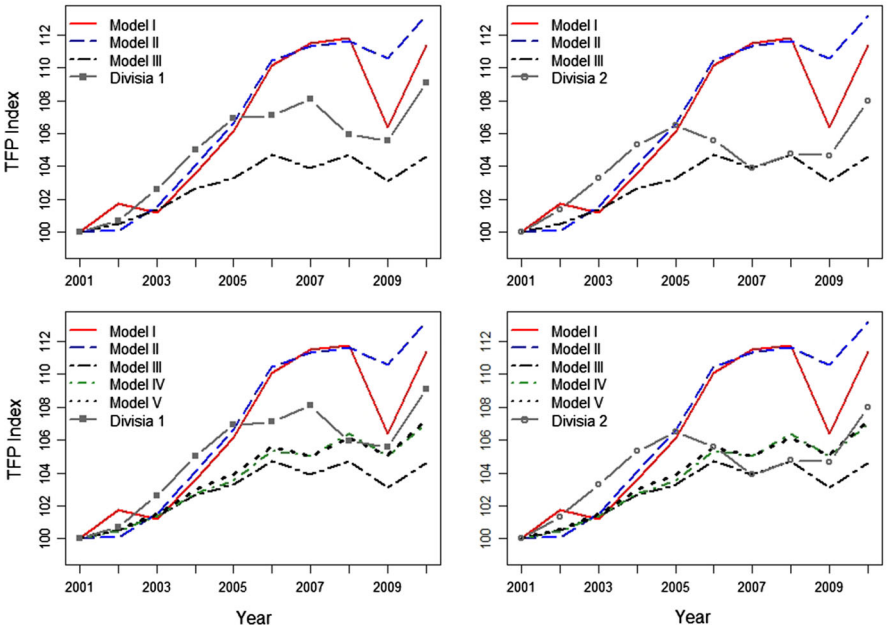

Fig. 4 TFP indices based on Models I–V

In this paper, we argue that the underlying production process in banking—namely, the production of (income-generating) credit—is inherently uncertain and show that the *ex-post* estimates of the banking technology (that assume uncertainty away) are likely to be biased and inconsistent. We offer an alternative methodology to recover banks' technologies based on the *ex-ante* cost function, which models credit uncertainty explicitly by recognizing that bank managers minimize costs subject to given *expected* outputs and credit risks. In order to make this model feasible to estimate, we estimate unobservable expected outputs and associated credit risk levels from banks' supply functions via nonparametric kernel methods. We apply this *ex-ante* framework to data on commercial banks operating in the U.S. during the period from 2001 to 2010.

We find that methods estimating the *ex-post* realization of banking technology tend to over-estimate output elasticities of cost which, in turn, leads to downward biases in the returns to scale estimates. The results, however, do not differ qualitatively across *ex-ante* and *ex-post* models if one controls for unobserved bank-specific effects. In this case, we find that virtually all U.S. commercial banks (regardless of the size) operate under IRS, which is consistent with findings recently reported in the literature despite the differences in methodology (e.g., Feng and Serletis 2010; Hughes and Mester 2013; Wheelock and Wilson 2012). Interestingly, when we leave bank-specific effects uncontrolled (as, for instance, the three above-cited studies do), the results change dramatically: the *ex-ante* models then indicate that 23–35 and 33–34 % of large banks exhibit decreasing and CRS, respectively.

Our empirical results show that the *ex-post* models tend to underestimate banks' TFP growth. For medium and large banks, the TFP growth estimates from *ex-ante* models tend to be higher than those from *ex-post* models. The opposite is true for

small banks. In fact, based on *ex-ante* models, we find that the average annual TFP growth is negative among small banks. All models suggest that the bulk of the positive productivity growth in the industry comes from the scale economies component. The *ex-ante* models estimate the asset-weighted average annual TFP growth due to IRS to be around 2.1–2.2 % per annum for banks of all sizes. Despite that small banks exhibit higher economies of scale, on average, the scale component in TFP is larger for medium and large banks. We find little evidence of economically significant technical progress, except for large banks. Risk level does not seem to impact productivity much either. We find no effect of expected risk levels on small and medium banks' productivity growth; little negative effect is found for large banks.

Acknowledgments Restrepo acknowledges financial support from the Colombian Fulbright Commission, the Colombian Administrative Department of Science, Technology and Innovation (Colciencias) and EAFIT University.

Appendix

See Table 6.

Table 6 Call report definitions of the variables

Variable	Call report definition	Description
y_1	rcfd1975	(Total issued) loans to individuals
y_1^+	y_1 − rcfd1979 − rcfd1981	y_1, less nonperforming loans to individuals
y_2	rcfd1410	(Total issued) real estate loans
y_2^+	y_2 − rcfd1422 − rcfd1423	y_2, less nonperforming real estate loans
y_3	rcfd1766 + rcfd1590 + rcfd3484 + rcfd3381 + rcfd2081 + rcfd1288 + rcfd2107 + rcfd1563	(Total issued) commercial and industrial loans, loans to finance agricultural production and other loans to farmers, lease financing receivables, interest-bearing balances due from depository institutions, loans to foreign governments and official institutions, loans to depository institutions, obligations (other than securities and leases) of states and political subdivisions in the U.S., other loans
y_3^+	y_3 − rcfd1583 − rcfd1607 − rcfd1608 − rcfd1227 − rcfd1228 − rcfd5381 − rcfd5382 − rcfd5390 − rcfd5391 − rcfd5460 − rcfd5461	y_3, less nonperforming categories that enter y_3
y_4	rcfd3365 + rcfd3545 + rcfd1754 + rcfd1773	(Total issued) federal funds sold and securities purchased under agreements to resell, trading assets, held-to-maturity securities total, available-for-sale securities
y_4^+	y_4 − rcfd3506 − rcfd3507	y_4, less nonperforming categories that enter y_4

Table 6 continued

Variable	Call report definition	Description
x_1	riad4150	Number of full-time equivalent employees on payroll at end of current period
x_2	rcfd2145	Premises and fixed assets
x_3	rcfd3353 + rcfd3548 + rcfd3190 + rcfd3200 + rcon2604	All borrowed money
x_4	rcon3485	Interest-bearing transaction accounts
x_5	rcfd2200 − rcon3485 − rcon2604	Non-transaction accounts: total deposits, less interest-bearing transaction accounts, less time deposits of $100,000 or more
w_1	riad4135/x_1	Salaries and employee benefits, divided by x_1
w_2	riad4217/x_2	Expenses on premises and fixed assets, divided by x_2
w_3	(riad4180 + riad4185 + riad4200 + riada517)/x_3	Expense of federal funds purchased and securities sold under agreements to repurchase, interest on trading liabilities and other borrowed money, interest on notes and debentures subordinated to deposits, interest on time deposits of $100,000 or more, divided by x_3
w_4	riad4508/x_4	Interest on transaction accounts (now accounts, ats accounts, and telephone and preauthorized transfer accounts) , divided by x_4
w_5	(riad4170 − riad4508 − riada517)/x_5	Interest on deposits, less interest on transaction accounts, less interest on time deposits of $100,000 or more, divided by x_5
p_1	riad4013/y_1^+	Interest and fee income on loans to individuals for household, family and other personal expenditures, divided by y_1^+
p_2	riad4011/y_2^+	Interest and fee income on loans secured by real estate, divided by y_2^+
p_3	(riad4012 + riad4024 + riad4065 + riad4115 + riad4056 + riad4058)/y_3^+	Interest and fee income on commercial and industrial loans, interest and fee income on loans to finance agricultural production and other loans to farmers in domestic offices, interest income on balances due from depository institutions, income from lease financing receivables, interest income on balances due from depository institutions, interest and fee income on all other loans in domestic offices, divided by y_3^+

Table 6 continued

Variable	Call report definition	Description
p_4	(riad4020 + riad4069 + riada220 + riad4218 + riad3521 + riad3196)/y_4^+	Interest income on federal funds sold and securities purchased under agreements to resell, interest income from trading assets, trading revenue, interest and dividend income on securities, realized gains (losses) on held-to-maturity securities, realized gains (losses) on held-to-maturity securities, realized gains (losses) on available-for-sale securities, divided by y_4^+
C	$\sum_{j=1}^{5} x_j w_j$	Total variable cost
inc	riad4079 − riad4080	Net noninterest income, less service charges on deposits
k	*quarterly average* of riad3210	Quarterly average of equity
Assets	*quarterly average* of rcfd2170	Quarterly average of total assets

References

Antle JM (1983) Testing the stochastic structure of production: a flexible moment-based approach. J Bus Econ Stat 1(3):192–201

Berger AN, Humphrey DB (1997) Efficiency of financial institutions: international survey and directions for future research. Eur J Oper Res 98(2):175–212

Berger AN, Mester LJ (1997) Inside the black box: what explains differences in the efficiencies of financial institutions? J Bank Finance 21(7):895–947

Berger AN, Mester LJ (2003) Explaining the dramatic changes in performance of US banks: technological change, deregulation, and dynamic changes in competition. J Financial Intermed 12(1):57–95

Berger AN, Hanweck GA, Humphrey DB (1987) Competitive viability in banking: scale, scope, and product mix economies. J Monet Econ 20(3):501–520

Chavas JP (2004) Risk analysis in theory and practice. Elsevier, San Diego

Clark JA (1996) Economic cost, scale efficiency, and competitive viability in banking. J Money Credit Bank 28(3):342–364

Deaton A, Muellbauer J (1980) An almost ideal demand system. Am Econ Rev 70:312–326

Denny M, Fussy M, Waverman L (1981) The measurement and interpretation of total factor productivity in regulated industries, with an application to Canadian Telecommunications. In: Cowing TG, Stevenson RE (eds) Productivity measurement in regulated industries, chap 8. Academic Press, New York

Feldstein MS (1971) Production with uncertain technology: some economic and econometric implications. Int Econ Rev 12:27–38

Feng G, Serletis A (2009) Efficiency and productivity of the US banking industry, 1998–2005: evidence from the Fourier cost function satisfying global regularity conditions. J Appl Econom 24(1):105–138

Feng G, 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. J Bank Finance 34(1):127–138

Feng G, Zhang X (2012) Productivity and efficiency at large and community banks in the US: a Bayesian true random effects stochastic distance frontier analysis. J Bank Finance 36(7):1883–1895

Freixas X, Rochet J (2008) Microeconomics of Banking. MIT Press, Cambridge

Hughes JP, Mester LJ (1993) A quality and risk-adjusted cost function for banks: evidence on the "too-big-to-fail doctrine. J Prod Anal 4(3):293–315

Hughes JP, Mester LJ (1998) Bank capitalization and cost: evidence of scale economies in risk management and signaling. Rev Econ Stat 80(2):314–325

Hughes JP, Mester LJ (2013) Who said large banks don't experience scale economies? Evidence from a risk-return-driven cost function. J Financial Intermed 22(4):559–585
Hughes JP, Lang W, Mester LJ, Moon CG (1996) Efficient banking under interstate branching. J Money Credit Bank 28(4):1045–1071
Hughes JP, Lang W, Mester LJ, Moon CG (2000) Recovering risky technologies using the almost ideal demand system: an application to US banking. J Financial Serv Res 18(1):5–27
Hughes JP, Mester LJ, Moon CG (2001) Are scale economies in banking elusive or illusive?: Evidence obtained by incorporating capital structure and risk-taking into models of bank production. J Bank Finance 25(12):2169–2208
Just RE, Pope RD (1978) Stochastic specification of production functions and econometric implications. J Econom 7:67–86
Just RE, Pope RD (eds) (2002) A comprehensive assessment of the role of risk in US agriculture. Kluwer, Norwell
Kumbhakar SC, Lovell CAK (2000) Stochastic frontier analysis. Cambridge University Press, Cambridge
Kumbhakar SC, Lozano-Vivas A (2005) Deregulation and productivity: the case of Spanish banks. J Regul Econ 27:331–351
Li Q, Racine J (2004) Cross-validated local linear nonparametric regression. Statistica Sinica 14(2):485–512
McAllister PH, McManus D (1993) Resolving the scale efficiency puzzle in banking. J Bank Finance 17(2):389–405
Moschini G (2001) Production risk and the estimation of ex-ante cost functions. J Econom 100(2):357–380
Pope RD, Chavas JP (1994) Cost functions under production uncertainty. Am J Agric Econ 76(2):196–204
Pope RD, Just RE (1996) Empirical implementation of ex ante cost functions. J Econom 72(1):231–249
Pope RD, Just RE (1998) Cost function estimation under risk aversion. Am J Agric Econ 80(2):296–302
Racine J, Li Q (2004) Nonparametric estimation of regression functions with both categorical and continuous data. J Econom 119(1):99–130
Restrepo-Tobón DA, Kumbhakar SC, Sun K (2013) Are US commercial banks too big? Working Paper, Binghamton University
Sealey CW, Lindley JT (1977) Inputs, outputs, and a theory of production and cost at depository financial institutions. J Finance 32(4):1251–1266
Shen CH, Chen YK, Kao LF, Yeh CY (2009) Bank liquidity risk and performance. Working Paper
Solow RM (1957) Technical change and the aggregate production function. Rev Econ Stat 39(3):312–320
Wheelock DC, Wilson PW (2001) New evidence on returns to scale and product mix among US commercial banks. J Monet Econ 47(3):653–674
Wheelock DC, Wilson PW (2012) Do large banks have lower costs? New estimates of returns to scale for US banks. J Money Credit Bank 44(1):171–199