J Regul Econ (2015) 48:125–168
DOI 10.1007/s11149-015-9275-1

ORIGINAL ARTICLE

Obelix vs. Asterix: Size of US commercial banks and its regulatory challenge

Diego Restrepo-Tobón · Subal C. Kumbhakar · Kai Sun

Published online: 1 April 2015

Abstract Big banks pose substantial costs to society in the form of increased systemic risk and government bailouts during crises. So the question is: Should regulators limit the size of banks? To answer this question, regulators need to assess the potential costs of such regulations. If big banks enjoy substantial scale economies (i.e., average costs get lower as bank size increases), limiting the size of banks through regulations may be inefficient and likely to reduce social welfare. However, the literature offers conflicting results regarding the existence of economies of scale for the biggest US banks. We contribute to this literature using a novel approach to estimating nonparametric measures of scale economies and total factor productivity (TFP) growth. For US commercial banks, we find that around 73 % of the top one hundred banks, 98 % of medium and small banks, and seven of the top ten biggest banks by asset size exhibit substantial economies of scale. Likewise, we find that scale economies contribute positively and significantly to their TFP growth. The existence of substantial scale economies raises an important challenge for regulators to pursue size limit regulations.

D. Restrepo-Tobón (✉)
EAFIT University, Carrera 49 #7 Sur 50, Medellín, Colombia
e-mail: drestr16@eafit.edu.co

S. C. Kumbhakar
Binghamton University, 4400 Vestal Pkwy E, Binghamton, NY 13902, US
e-mail: kkar@binghamton.edu

S. C. Kumbhakar
University of Stavanger Business School, 4036 Stavanger, Norway

K. Sun
University of Salford, Salford, Greater Manchester M5 4NT, UK
e-mail: ksun1@binghamton.edu

Keywords Bank regulation · Economies of scale · Returns to scale · Nonparametric methods

JEL Classification L51 · G21 · G28 · D24 · L25 · C14

1 Introduction

After the latest US financial crisis, the optimal size of banks became an important issue for regulators and policymakers. Too-big-to-fail banks seems to enjoy a funding advantage in capital markets over smaller banks mostly due to implicit government guarantees (Gandhi and Lustig, In press). Prominent academics and policymakers have proposed to break up the biggest banks arguing that they pose substantial costs to society in the form of government bailouts and increased systemic risk.[1] However, policy proposals to break up or to cap the size of the biggest US banks failed to materialize. A plausible explanation for this outcome is that there is no consensus on the desirability or feasibility of capping the size of banks (Stern and Feldman 2009). More recently, the discussion has been reignited and capping the size of banks is again at the center of the regulatory policy debate.

Against this backdrop, the existence of economies of scale for the biggest US banks has become a contentious and important issue. Economies of scale (which imply reduction in average costs as firm's size increases), have been at the center of policy analyses in the regulation/deregulation and reregulation of different industries.[2] It would be challenging for regulators to promote perfect competition when substantial economies of scale are present (Laudati 1981). As suggested by Borts (1954), the existence of scale economies may lead regulators to favor the emergence and consolidation of large firms while ensuring that customers are served on reasonable terms—even if new entry and competition in the market is reduced.

Since Williamson (1968, 1977) economies of scale underlie almost every aspect of antitrust regulation and constitute the basis for the well-known *efficiency defense* in mergers.[3] For instance, a merger that lowers competition may bypass antitrust restrictions if the merging parties can demonstrate that the merger is necessary to achieve substantial efficiencies that will enhance consumers' welfare. DeYoung (1991) shows that the Williamsonian efficiency defense was implemented by the Federal Reserve Board of Governors in deciding over bank mergers even though the US Supreme Court had not recognized offsetting efficiencies as a defense of an anticompetitive merger.[4]

[1] See Johnson (2012) and the references therein.

[2] See for instance Davies and Tracey (2014), Glass and Stefanova (2012), Camacho and Menezes (2009), Nauges and Van Den Berg (2008), Kumbhakar and Wang (2007), Evans and Guthrie (2006), Fraquelli et al. (2005), Foreman and Beauvais (1999), Ellig and Giberson (1993), and Evanoff et al. (1990).

[3] See Williamson (1968), Berger and Humphrey (1992), Farrell and Shapiro (2001), Schmalensee (2004), McAfee et al. (2004), Lagerlof and Heidhues (2005).

[4] See Kolasky and Dick (2003) and Kinne (1998) for extensive analyses of how the *efficiency defense* evolved in the *US Merger Guidelines and Laws* and some merger examples in which scale economies considerations were taken into account.

Conventional wisdom says that as firms get bigger economies of scale tend to shrink towards zero–ultimately reaching a point where the efficient scale size is achieved. Thus, regulating their size is unnecessary. However, if banks enjoy scale economies, regulators and policymakers have to balance the benefits with the costs of further bank growth. Lower average costs may allow banks to offer more competitive prices on their services, benefiting consumers and society as a whole. On the other hand, large scale banks may increase systemic risk and crisis costs. Thus, determining if scale economies exist for banks has important implications for bank regulation (Feldman 2010).

The literature on the existence of scale economies for US banks is inconclusive. Early studies find scale economies for small banks only. According to Berger and Humphrey (1994), the conclusion that scale economies were available only to smaller US banks was unshaken at the time. In contrast, more recent studies find scale economies even for the biggest financial institutions (e.g. Wheelock and Wilson 2012; Hughes and Mester 2013; Malikov et al. 2014). However, other studies find no evidence of either scale economies or increasing returns to scale (RTS) for large US commercial banks (e.g. Feng and Zhang 2012; Restrepo-Tobón and Kumbhakar 2014; Davies and Tracey 2014). Thus, the existence of scale economies for the biggest US commercial banks is still a matter of debate.

We contribute to this literature by measuring scale economies using a novel nonparametric approach that avoids making restrictive assumptions regarding the functional form of the underlying technology for banks (see Wheelock and Wilson 2011, 2012). Our approach offers two important advantages over the nonparametric methods used in previous studies. First, it gives fully nonparametric estimates of scale economies and total factor productivity (TFP) growth and its components. Previous nonparametric studies did not examine the relation between scale economies and TFP growth and did not allow estimation including some important features of the underlying technology. Many studies did not report measures like cost elasticities of outputs, technical change (TC), cost elasticities of input prices, among others, making it difficult to check the consistency of their results with economic theory. Second, unlike previous studies, our method allows us to partially control for unobserved heterogeneity across banks.[5]

[5] We model the unobserved heterogeneity across banks as additive fixed-effects. This is the standard approach to control for unobserved heterogeneity in parametric models (see Chamberlain 1984; Arellano and Honore 2001, and Bai 2009). Ignoring additive unobserved heterogeneity might lead to bias and inconsistency (see Mester 1997; Greene 2005a, b; Kumbhakar et al. 2008, and Wang and Ho 2010). However, input prices and outputs can vary systematically as a function of input characteristics and bank business models, which may vary across banks. So input prices and outputs can have different meaning across different banks, and not all of these variations will necessarily be captured in the bank fixed-effects. Further, unobserved heterogeneity can also be due to differences in costly risk-taking among banks. Note that in addition to controlling for additive fixed bank-specific effects, in our preferred model we allow the "core" technology to be nonparametric which allows full flexibility regarding the marginal effects of input prices and outputs. Since these effects are both bank and year-specific, differences in input prices and outputs among banks are fully captured by the nonparametric specification of the technology. In addition, we control for output quality (through nonperfoming loans), risk (via equity), and nontraditional activities (via noninterest income) to capture heterogeneity across banks that may be confounded with differing relative costs (e.g. Mester 1996; Rogers 1998) or masquerade sources of risk (e.g. Hughes and Mester 1998).

Our empirical findings indicate that failing to control for unobserved heterogeneity may conceal evidence of scale economies.

Overall, our findings suggest that most US commercial banks with assets in excess of $1 billion enjoy substantial economies of scale (Obelix is not obese).[6] However, 35 % of the observations for the top one hundred banks show no evidence of scale economies.[7] For the top ten banks with assets ranging from $47 billion to $1.5 trillion, only 70 % of the observations are consistent with economies of scale.[8] In particular, of the four banks with assets above $500 billion, on average, only one bank exhibits economies of scale during the sampling period (only 3 obese Obelix). Thus, capping the size of the biggest banks (converting them to Asterix) may yield limited social losses from the scale economies viewpoint.

The rest of the paper is organized as follows. In Sect. 2, we review the literature. In Sect. 3, we describe our model, the estimation strategy, and our data. In Sect. 4, we present our empirical results and compare them with those in the existing literature. Section 6 discusses robustness checks. Policy implications of our study are discussed in Sect. 6. Finally, Sect. 7 concludes.

2 Related literature

During the early stages of deregulation of the US banking industry, regulators and policymakers were concerned that a handful of big diversified banks could emerge and dominate the industry. Banks appeared to be able to profit from being bigger, threatening the viability of smaller specialized banks (Clark 1988). More recently, empirical evidence shows that advances in technology and changes in regulation favor large depository institutions (Wheelock and Wilson 2011). Economists think that economies of scale may explain such phenomena. However, the empirical evidence seems, at best, inconclusive.

Before 1970, empirical studies showed evidence of scale economies among commercial banks of all sizes (see Shaffer 1994 and the references therein). During the 1970's and 1980's the bulk of empirical evidence, in contrast, showed scale economies for only small depository institutions.[9] In the 1990's, the evidence was mixed: some studies find evidence of scale economies for small banks only[10] while others present

[6] Following the literature, we classify big banks as those with assets above $1 billion dollars; medium banks as those with assets greater than $100 million and lower than $1 billion dollars; and small banks as those with assets below $100 million dollars. For each year, we also rank banks by assets and select the top 100 biggest banks. As of December 31, 2010; there were 479 big banks; 3637 medium banks; and 2290 small banks operating in the US. By comparison, as of December 31, 2000, there were 285 big banks; 2397 medium banks; and 5452 small banks. All nominal variables are in 2005 dollars.

[7] The banks belonging to the top one hundred banks have assets ranging from $23 million to $1.5 million.

[8] The top ten biggest banks for each year are: State Street Bank and Trust Company, CitiBank, US Bank, Wachovia, HSBC, Wells Fargo, Bank of America, Bank of New York Mellon, Fleet National Bank, Suntrust Bank, Keybank, PNC Bank, Regions Bank, JP Morgan Chase, and Citizens Bank.

[9] See Clark (1988), Berger et al. (1987), and Shaffer (1994).

[10] See Berger and Humphrey (1994), Berger et al. (1987), Hunter and Timme (1991), Mester (1994), Clark (1996), Hughes and Mester (1998), and Wheelock and Wilson (2001).

evidence of scale economies even for the largest banks.[11] More recent studies show evidence of scale economies across the entire bank size distribution. These studies include Wheelock and Wilson (2011, 2012), Hughes and Mester (2013), and Feng and Serletis (2010). However, Feng and Zhang (2012) find no evidence consistent with economies of scale for large US commercial banks; Restrepo-Tobón and Kumbhakar (2014) find no evidence of significant economies of scale for large US banks and bank holding companies; and Davies and Tracey (2014), after controlling for Too-big-to-fail factors, find no evidence of scale economies for a sample of large US banks.

Wheelock and Wilson (2011, 2012) investigate scale economies of US credit unions, commercial banks, and bank holding companies. Despite the continuous growth of these institutions since the 1980's, they uncovered substantial economies of scale even for the largest institutions. Using yearly data from 1989–2006, Wheelock and Wilson (2011) find evidence consistent with economies of scale for all US credit unions and conclude that further consolidation of the industry and increasing average size are likely. Regarding commercial banks and bank holding companies, Wheelock and Wilson (2012) find similar results. Using quarterly data from 1984–2006, their estimates support the existence of economies of scale for 99.7 % of the observations. These results contrast with their earlier work, viz., Wheelock and Wilson (2001), which shows that economies of scale were exhausted when banks have assets between $300 and $500 million. They attribute their contrasting results to the use of a larger sample and a more realistic model of bank costs in their latest work in which equity capital and off-balance sheet activities are explicitly incorporated.

Hughes and Mester (2013) present evidence of economies of scale for large US bank holding companies. They use parametric methods to estimate a model accounting for managerial risk preferences and the endogenous risk taking behavior of banks. Large banks may have lower marginal costs in risk management due to diversification. This, in turn, may lead banks to take on more risk until all the scale-related cost savings are exhausted. Thus, failing to account for managerial preferences for risk and endogenous risk taking might lead to biased measures of economies of scale.[12]

Hughes and Mester (2013) use four different model specifications. The first model omits equity capital as a conditioning variable and also omits the cost of equity in the cost function. Using this model, they find little economies of scale for all bank holding companies in the sample. Using a second model, in which equity capital enters as a conditioning variable, they find diseconomies of scale for all bank holding companies. Results from a third model incorporating the shadow cost of equity shows evidence of slight economies of scale for all bank holding companies. However, the estimates of economies of scale are statistically indistinguishable from those of the first model. Finally, using a fourth model incorporating managerial preferences and conditioning on the optimal level of equity capital, they find substantial scale economies for all bank holding companies. In this case, estimates of economies of scale range from 1.13 for

[11] See Shaffer (1994), Hughes et al. (1996, 2000), Berger and Mester (1997), Hughes and Mester (1998), Berger et al. (1999), and Hughes et al. (2001).

[12] See also Hughes et al. (1996), Hughes and Mester (1998), Hughes et al. (2001), and Hughes and Mester (2010).

the smallest bank holding companies to 1.37 for the largest. Contrary to standard economic theory, however, Hughes and Mester (2013)'s economies of scale estimates increase monotonically with bank holding companies' size. This result is puzzling since most empirical evidence, including Wheelock and Wilson (2011, 2012), show that economies of scale estimates decrease as bank size increases.

Feng and Serletis (2010) find evidence of increasing returns to scale (RTS) for large banks. Unlike Wheelock and Wilson (2011, 2012) and Hughes and Mester (2013) who use a cost function, Feng and Serletis (2010) use a fully parametric output distance function which does not require data on input prices. They use data for US banks with assets in excess of $1 billion from 2001 to 2005. The main contribution of their paper is methodological in nature. They impose monotonicity and curvature constraints on the underlying bank production technology using Bayesian techniques. They claim that without such constraints, estimates of RTS and TFP growth are distorted. They obtain RTS estimates ranging from 1.037 to 1.056, indicating moderate increasing RTS for all banks in the sample. Nonetheless, from their paper, it is not clear if these results are entirely driven by their method, since they omit a comparison with the alleged misspecified model. Furthermore, in almost all banking studies, outputs are treated as exogenous and inputs as endogenous. If this is the case, one should use an input distance function in which outputs and input ratios appearing as regressors are exogenous (Das and Kumbhakar 2012). The use of output distance functions suffers from endogeneity problem when inputs are endogenous.

Malikov et al. (2014) estimate banks' production technology based on the ex-ante cost function. They model credit uncertainty explicitly by recognizing that bank managers minimize costs subject to given expected outputs and credit risk. They find that virtually all US commercial banks (regardless of the size) operate with economies of scale. They also show that failing to control for unobserved heterogeneity across banks may conceal evidence of economies of scale and that methods estimating the ex-post realization of banking technology lead to downward biases in economies of scale estimates.

Taken together, these studies provide evidence of economies of scale for most banks. However, in a recent paper, Feng and Zhang (2012) failed to reject constant or decreasing RTS for large and small banks. They use an output distance function to model banks' technology and Bayesian techniques for estimation. Further, they impose monotonicity and curvature conditions to the underlying technology. They include continuously operating large banks (assets above $1 billion), large community banks (assets below $1 billion and above $100 million), and small community banks (assets below $100 million) from 1997 to 2006. The main distinction with Feng and Serletis (2010) is that Feng and Zhang (2012) explicitly account for technical inefficiency and random unobserved heterogeneity across banks in their estimation. Their results suggest that failing to do so leads to higher RTS estimates. In particular, without accounting for unobserved heterogeneity, large banks exhibit increasing RTS ranging from 1.022 in 1997 to 1.01 in 2006. However, when unobserved heterogeneity is incorporated these results disappear: all large banks now exhibit constant RTS.

More recently, Restrepo-Tobón and Kumbhakar (2014) find no evidence of significant increasing RTS for large US banks and bank holding companies. They use a nonparametric approach based on input distance functions which need no information

on input prices, partially controlling for too-big-to-fail cost advantages embedded in input prices. Their results show that although some big banks are operating under increasing RTS, estimates of RTS are found to be numerically close to unity. Davies and Tracey (2014), using a parametric approach and explicitly controlling for too-big-to-fail factors, also find no evidence of scale economies for a sample of large US banks. Therefore, the literature is still inconclusive regarding the existence of economies of scale for the biggest banking organizations in the US.

Thus, the literature will benefit from further investigation on scale economies and TFP growth using a methodology that (i) controls for unobserved heterogeneity across banks, (ii) accounts for bank risk taking, and (iii) uses flexible functional forms to model the underlying technology.[13] Our paper goes one step forward in this direction.[14]

Another way to rationalize the observed increasing size of banks is to look at the distribution of TFP components across different bank size categories. If TFP growth differs across the bank size distribution, some banks will have incentives to increase their scale of operations, even though, by itself, this strategy may have negligible or negative impact on their performance. For instance, Hunter and Timme (1986, 1991) investigate how TC interacts with bank size and scale economies. Using data for US bank holding companies from 1972 to 1982, they find that TC is associated with an increase in scale economies: technological change lowers costs by 1 % per year, increases the cost-minimizing scale of operations, and affects the product mix of banks.

Early studies indicate that TFP growth in the US banking industry during the 1980's was modest or slightly negative (Humphrey 1991, 1992, 1993; Bauer et al. 1993; Wheelock and Wilson 1999; Stiroh 2000; Semenick Alam 2001; Berger and Mester 2003; Daniels and Tirtiroglu 1998). Tirtiroglu et al. (2005) investigate how the regulatory structure at the time might have contributed to this phenomenon. They find that deregulation had a positive impact on banks' long-run productivity growth, confirming the belief that the poor productivity growth was due to regulatory restrictions. Mukherjee et al. (2001) and Semenick Alam (2001) find positive productivity growth for large US banks over roughly the same period, indicating that larger banks may have taken better advantage of technological progress and deregulation.

Using data from 1997 to 2006, Feng and Zhang (2012) find positive productivity growth for large banks, 2.04 % per year on average, but poor productivity growth for small banks, 0.3 % per year on average. They also show that most productivity growth gains are due to TC and not to efficiency gains.[15] These results echo those of Feng and Serletis (2010) who show that technical change contributed to 70 % of total productivity growth and scale economies account for only 7 %.[16]

[13] The international evidence also points toward the existence of scale economies only for small and medium-sized banks (Amel et al. 2004). Allen and Liu (2007), on the other hand, report economies of scale for the six largest Canadian banks.

[14] Restrepo-Tobón and Kumbhakar (2013) using data for all US commercial banks from 2001 to 2010 also document constant RTS for most banks in their sample.

[15] The results for large banks are comparable to those in Feng and Serletis (2010).

[16] See Hughes et al. (2001) and Hughes and Mester (2010) on the impact of using different models on estimates of TFP components (e.g., scale economies, efficiency, technical change, etc.).

3 Methodology

We avoid the problem of specifying *a priori* any functional form for the underlying cost function. Unlike previous studies, our method exploits the dynamics between total cost and its determinants over time. We start from a purely nonparametric specification of the cost function for banks and then derive a growth equation for costs whose coefficients are nonparametric functions of the arguments of the cost function, viz., outputs, input prices, and other environmental/control variables. Similarly to Kumbhakar and Sun (2012), we use the semiparametric smooth coefficient (SPSC) model of Li et al. (2002) and Li and Racine (2010) to estimate the functional coefficients. The SPSC model arises naturally from the cost function framework and is not imposed *a priori*. Moreover, the functional coefficients of the SPSC model are nonparametric functions of the covariates of the cost function and are related to TFP growth and its components, including economies of scale.

3.1 The model

We use a nonparametric cost function with Q outputs, K inputs, and P environmental/control variables to account for some bank-specific characteristics. The cost function for bank i at time t has the following general specification:

$$C_{it} = A_{it} \cdot B_i \cdot H\left(W_{it}, Y_{it}, t, Z_{it}\right) \tag{1}$$

where C_{it} represents actual cost, A_{it} is a productivity parameter, B_i is a parameter capturing time-invariant unobserved heterogeneity across banks (fixed-effects), $H(\cdot)$ is a nonparametric function of W_{it}, a vector of input prices, Y_{it}, a vector of output quantities, and Z_{it}, a vector of control variables.

We impose linear homogeneity restrictions on input prices by dividing costs and input prices by the price of input K, i.e., W_K. Thus, the cost function becomes:

$$\widetilde{C}_{it} = A_{it} \cdot B_i \cdot H\left(\widetilde{W}_{it}, Y_{it}, t, Z_{it}\right) \tag{2}$$

where $\widetilde{W}_k = W_k / W_K$ and $\widetilde{C}_{it} = C / W_K$. Taking logarithm to Eq. (2) and denoting $\ln H(\cdot) = f(\cdot)$, we can write:

$$\ln \widetilde{C}_{it} = f\left(\widetilde{W}_{it}, Y_{it}, t, Z_{it}\right) + \ln A_{it} + \ln B_i. \tag{3}$$

Dropping the subscripts and taking the total differential of Eq. (3), we get the following growth formulation:

$$\frac{d \ln \widetilde{C}}{dt} = \frac{\partial f}{\partial t} + \sum_{k=1}^{K-1} \frac{\partial f}{\partial \ln \widetilde{W}_k} \cdot \frac{d \ln \widetilde{W}_k}{dt} + \sum_{q=1}^{Q} \frac{\partial f}{\partial \ln Y_q} \cdot \frac{d \ln Y_q}{dt} + \sum_{p=1}^{P} \frac{\partial f}{\partial Z_p} \cdot \frac{dZ_p}{dt} + \frac{\partial \ln A}{\partial t}. \tag{4}$$

We write Eq. (4) in a more compact way to get our estimating growth equation:

$$\dot{\tilde{C}} = \beta_0(\cdot) + \sum_{k=1}^{K-1} \beta_k(\cdot)\dot{\tilde{W}}_k + \sum_{q=1}^{Q} \gamma_q(\cdot)\dot{Y}_q + \sum_{p=1}^{P} \varphi_p(\cdot)\nabla_t Z_p + u \tag{5}$$

where, in generic terms, $\dot{X} = d\ln X/dt$ and $\nabla_t X = dX/dt$. We interpret $u = \partial \ln A/\partial t$ as an error term capturing productivity shocks and establish the following mapping for the functional coefficients:

$$\beta_0(\cdot) = \frac{\partial f}{\partial t}; \quad \beta_k(\cdot) = \frac{\partial f}{\partial \ln \widetilde{W}_k}; \quad \gamma_q(\cdot) = \frac{\partial f}{\partial \ln Y_q}; \quad \varphi_p(\cdot) = \frac{\partial f}{\partial Z_p}. \tag{6}$$

Note that all the functional coefficients in Eq. (5), i.e., $\beta_0(\cdot)$, $\beta_k(\cdot)$, $\gamma_q(\cdot)$, and $\varphi_p(\cdot)$ are functions of $\widetilde{W}_{it}, Y_{it}, t, Z_{it}$. Further, the unobserved time-invariant heterogeneity across banks (fixed-effects), represented by the parameters $\ln B_i$ in Eq. (3), are removed in the growth formulation of the cost function in Eq. (5). Finally, the functional coefficients have clear economic meaning. This is shown by relating them to TFP growth and its components in the next subsection.

3.2 Total factor productivity change

We start with the standard definition of TFP change (see Denny et al. 1979):

$$T\dot{F}P \equiv \sum_{q=1}^{Q} R_q \dot{Y}_q - \sum_{k=1}^{K} S_k \dot{X}_k \tag{7}$$

where R_q denotes the revenue share of each output ($q = 1, \ldots, Q$) and S_k the cost share of each input ($k = 1, \ldots, K$). In Appendix 1 we show that TFP growth can be expressed as:

$$T\dot{F}P \equiv -\beta_0(\cdot) + \sum_{q=1}^{Q}(R_q - \gamma_q(\cdot))\dot{Y}_q + \sum_{k=1}^{K-1}(S_k - \beta_k(\cdot))\,\dot{\tilde{W}}_k - \sum_{p=1}^{P}\varphi_p(\cdot)\nabla_t Z_p - u \tag{8}$$

The first term is the TFP growth's technical change component ($TC = -\partial f/\partial t = -\beta_0(\cdot)$) which captures shift of the cost function over time (percentage change in cost over time, *ceteris paribus*). $TC < 0$ indicates technical progress (cost diminution). A positive value of TC will indicate technical regress. The second term $\left(SC = \sum_{q=1}^{Q}(R_q - \gamma_q(\cdot))\dot{Y}_q\right)$ is the scale component which can be decomposed into two components by expressing it as $(RTS - 1)\sum_{q=1}^{Q}\gamma_q(\cdot)\dot{Y}_q + \sum_{q=1}^{Q}\left(R_q - \gamma_q(\cdot)/\Gamma(\cdot)\right)\dot{Y}_q$ where $\Gamma(\cdot) = \sum_q \gamma_q(\cdot)$ and $RTS = [\Gamma(\cdot)]^{-1}$. The first

part of it clearly depends on RTS and the second part depends on mark-up (departure of output prices from their respective marginal costs).

The third term $\left(AL = \sum_{k=1}^{K-1}(S_k - \beta_k(\cdot))\dot{\tilde{W}}_k\right)$ corresponds to the allocative component because it captures the effects of non-optimal input allocation (deviation of input mix from the optimal). The fourth term $\left(EX = -\sum_{p=1}^{P}\varphi_p(\cdot)\nabla_t Z_p\right)$ is the exogenous component which captures the effect of factors such as output quality, risk, and other variables. Finally, the last term $(-u)$ is viewed as a productivity shock component which can increase/decrease TFP growth. We assume it to be random with mean zero and can be measured residually. Note that the TFP growth components are nonparametric functions and each of these components can be computed after estimating Eq. (5).

3.3 Econometric model

We estimate the functional coefficients in Eq. (5) using three different models. First, note that $f(\cdot)$ in Eq. (5) can be thought of as an unknown smooth function of t, $\ln \widetilde{W}_k$, $\ln Y_q$, and Z_p and its gradients, i.e., $\beta_0, \beta_k,\ \forall k = 1, \ldots, K-1; \gamma_q,\ \forall q = 1, \ldots, Q$ and $\varphi_p,\ \forall p = 1, \ldots, P$ are also unknown smooth functions of these variables as shown in Eq. (6). Thus, Eq. (5) can be viewed as the SPSC model of Li et al. (2002) where the model is linear in $\dot{\tilde{W}}_k,\ \forall k = 1, \ldots, K-1; \dot{Y}_q,\ \forall q = 1, \ldots, Q$ and $\nabla_t Z_p,\ \forall p = 1, \ldots, P$. To simplify notation, we rewrite Eq. (5) (after adding back the subscript i and t for bank and year, respectively) as

$$\mathcal{Y}_{it} = \mathcal{X}'_{it}\Psi(\mathcal{Z}_{it}) + u_{it} \tag{9}$$

where $\mathcal{Y}_{it} = \dot{\tilde{C}}_{it}; \mathcal{X}'_{it} = \left[1, \dot{\tilde{W}}_{1it}, \ldots, \dot{\tilde{W}}_{K-1\,it}, \dot{Y}_{1it}, \ldots, \dot{Y}_{Qit}, \nabla_t Z_{1it}, \ldots, \nabla_t Z_{Pit}\right]$, $\mathcal{Z}'_{it} = \left[\ln \widetilde{W}_{1it}, \ldots, \ln \widetilde{W}_{K-1\,it}, \ln Y_{1it}, \ldots, \ln Y_{Qit}, t_{it}, Z_{1it}, \ldots, Z_{Pit}\right]$ and $\Psi'(\cdot) = \left[\beta_0(\cdot), \beta_1(\cdot), \ldots, \beta_{K-1}(\cdot), \gamma_1(\cdot), \ldots, \gamma_Q(\cdot), \varphi_1(\cdot), \ldots, \varphi_P(\cdot)\right]$. We call this model the semiparametric growth model (**SPG model**). Estimation of this model is shown in details in Appendix 2.

Second, we estimate Eq. (5) assuming that the underlying technology can be represented by a translog cost function. In this case, the functional coefficients in Eq. (5) are parametric functions of $[W_{it}, Y_{it}, t, Z_{it}]$ that can be estimated by OLS (see Eq. (22) in Appendix 2). We call this model the parametric growth model (**PG model**).[17]

Third, to make our results comparable to the previous literature, we estimate the cost function Eq. (1) instead of Eq. (5) assuming that the underlying technology can be represented by a translog cost function. However, unlike the PG model, Eq. (1) does not control for unobserved time-invariant bank heterogeneity (see Appendix 2). We call this model the parametric log model (**PL model**). After estimating the coefficients

[17] Using both the SPG and PG models, we avoid the incidental parameter problem. Consistency of the functional coefficients follow from the standard SPSC model (Li et al. 2002) and the standard linear panel data models as the number of observations grow to infinity.

of the translog model, we compute the functional coefficients in Eq. (5) which are then used to compute TFP growth components in Eq. (8).

The three models described above (SPG, PG, and PL) differ in terms of their assumptions about the data generating process and the functional form of the underlying technology. Contrary to the PL model, both the SPG and PG models control for time-invariant unobserved bank heterogeneity. Unlike the SPG model, the PG model assumes a common parametric technology for all banks. The functional coefficients of the SPG model are fully nonparametric while in the the PG and PL models these functional coefficients are fully parametric. In this sense the SPG model is the most general among the three models.

3.4 Data

We focus on the post-deregulation period of the US banking industry to isolate economies of scale estimates from previous regulatory restrictions that may have blocked banks' ability to grow. We include the recent financial crisis to investigate how economies of scale of the biggest US banks were affected during this period. Our sample covers 60,868 bank-year observations for 7,473 commercial banks from 2001 to 2010. We exclude Bank Holding Companies (BHC) which, given their idiosyncrasies, may not be comparable with most US commercial banks. However, as a robustness check, we conduct a separate study regarding these institutions. Also, although quarterly data are available for balance-sheet figures, we do not believe that each quarterly observation reflects banks' economic behavior over the year. Hence, we take quarterly averages of balance-sheet figures to compute yearly observations.[18] Our approach incorporates a large sample of banks, a nonparametric specification of the cost function, fully flexible interactions among the explanatory variables, and includes off-balance sheet activities, nonperforming loans, and equity capital as additional control variables.[19]

We use data from the Report of Conditions and Income (Call Reports) from the Federal Reserve Bank of Chicago. We include all FDIC insured commercial banks with

[18] We use quarterly averages for balance-sheet figures only since end-of-year balance-sheet figures may differ from the average balance-sheet amounts a bank maintains over the year. For instance, if during the first three quarters a bank has $1 Billion in outstanding loans but this amount is reduced significantly in the fourth quarter to $800 million, the end-of-year amount will underestimate the total amount of loans that contributed to generate the reported interest income. In the case of balance-sheet figures used to estimate input prices, using end-of-year as opposed to quarterly averages, will lead to over-estimate them. To clarify, the approach used is to sum up the end-of-quarter figures within a specific year and divide the result by 4. We do not do this for income statement figures since end-of-year numbers are the figures we need.

[19] Risk complicates the analysis of a bank's production technology. The assumption of cost minimization may be inappropriate without accounting for risk because cost is affected by the degree of risk a bank takes. We accommodate risk indirectly by including some control variables like equity capital, nonperforming loans, and off-balance sheet activities. This is the standard approach in the literature. A more desirable approach would be to account for endogenous risk taking as in Hughes and Mester (2013) and Malikov et al. (2014). However, doing so will highly complicate our nonparametric estimation strategy. The results in Hughes and Mester (2013) and Malikov et al. (2014) suggest that the estimates of RTS are biased downwards when risk taking is not taken into account. Thus, our results can be viewed as the lower bounds of economies of scale.

available data between 2001Q1 and 2010Q4. We exclude banks reporting negative values for assets, equity, outputs and prices, stand alone internet banks, commercial banks conducting primarily credit card activities, and banks chartered outside continental US territory. Our data set is an unbalanced panel with 63,120 bank-year observations for 8,483 banks. We deflate all nominal quantities using the 2005 Consumer Price Index for all urban consumption published by the Bureau of Labor Statistics.

Our method is computationally demanding. To avoid extreme outliers, we use data for which output quantities and input prices fall between 0.5 and 99.5 % of their empirical distributions. In addition, we only consider banks with at least four years of available data from 2001 to 2010. To get annual values, we compute the quarterly average of balance-sheet nominal (stock) values. As a result, we have 60,868 bank-year observations for 7,473 different banks. Our growth formulation further decreases the number of observation to 51,966 bank-year observations for 7,381 banks. However, this growth formulation removes the bank-specific fixed-effects, and therefore it does not impose any extra cost in reducing degrees of freedom.

To estimate our model, we must map banks' activities to outputs and input quantities and their corresponding prices. We follow the literature and model bank activities using the balance-sheet approach of Sealey and Lindley (1977). In this framework, a bank's balance-sheet captures the essential structure of banks' core business: (i) liabilities, together with physical capital and labor, are inputs into the bank production process and (ii) assets, other than physical assets, are outputs. Liabilities are composed of core deposits and purchased funds. Assets include loans and trading securities. Therefore, banks use labor, physical capital, and debt to produce loans, invest in financial assets, and facilitate other financial services.

We define five output variables: household and individual loans (Y_1), real estate loans (Y_2), business loans (Y_3), securities (e.g., federal funds sold and securities purchased under agreements to resell) (Y_4), and other assets (Y_5). These outputs are essentially the same as those used in Berger and Mester (2003).

We define five input variables: labor (e.g., number of full-time equivalent employees at the end of each quarter) (X_1), physical capital (e.g. premises and fixed assets including capitalized leases) (X_2), purchased funds (e.g., time deposits of $100,000 or more, federal funds purchased and securities sold under agreements to repurchase, total trading liabilities, other borrowed money, and subordinated notes and debentures) (X_3), interest-bearing transaction accounts (X_4), and core deposits other than interest-bearing transaction accounts or time deposits of $100,000 or more ($X_5$). The price of each input, W_j for $j = 1, \ldots, 5$, is computed by dividing total expenses by the corresponding input quantity.[20] Total costs, C, equals the sum of expenses on each of the five inputs. Total revenue equals the sum of revenues for each output category. Table 1 presents summary statistics for outputs, inputs, and input prices.

[20] We treat physical capital as a variable input and compute its price for each year as the ratio between expenses of premises and fixed assets over the quarterly average of total premises and fixed assets. Ideally, one should use the opportunity cost of such assets for this computation. We believe, however, that since expenses of premises and fixed assets include all non-interest expenses related to the use of premises and fixed assets, including capitalized leases, net of their rental income, our approach may reasonably resemble the opportunity cost of using such assets.

Table 1 also presents information on cost, revenue, and output shares. On the revenue side, real estate loans account for 41.6 % of total banks' revenues (R_2), loan to business and other institutions 20.5 % (R_3), securities 15.3 % (R_4), and other assets 16.1 % (R_5). Loans to individuals and households account for 6 % (R_1) of total revenue.

Table 1 Summary statistics, 2001–2010

Variable	Mean	SD	Percentiles				
			5th	25th	50th	75th	95th
Y_1	70,358	1,700,000	81.50	655.7	4532	46,077	137,000,000
Y_2	360,000	5,390,000	1177	5683	49,638	546,000	404,000,000
Y_3	220,000	4,740,000	713.5	2953	17,131	166,000	348,000,000
Y_4	291,000	7,950,000	826.4	4601	27,012	247,000	782,000,000
Y_5	38,042	1,010,000	95.62	308.0	1595	17,445	89,500,000
X_1	226	3476	3.000	8.500	36.25	296.75	214,000
X_2	11,686	146,000	4.000	151.80	2022	19,613	10,500,000
X_3	290,000	6,670,000	24.628	2519	21,474	272,000	540,000,000
X_4	26,307	247,000	56.084	1757	10,096	53,636	20,000,000
X_5	545,000	11,000,000	1025	10,929	50,047	585,000	763,000,000
W_1	52.54	12.99	23.91	36.47	49.99	77.82	117.1
W_2	0.346	0.338	0.052	0.104	0.241	0.955	3.406
W_3	0.034	0.011	0.008	0.017	0.034	0.053	0.064
W_4	0.010	0.007	0.001	0.002	0.008	0.025	0.042
W_5	0.026	0.010	0.004	0.011	0.025	0.046	0.055
R	71,765	1,300,000	234	1792	8290	78,874	83,600,000
R_1	0.065	0.058	0.000	0.007	0.051	0.171	0.913
R_2	0.416	0.165	0.000	0.149	0.414	0.693	0.943
R_3	0.205	0.147	0.000	0.040	0.164	0.511	0.915
R_4	0.153	0.106	0.000	0.027	0.129	0.358	0.961
R_5	0.161	0.088	0.001	0.049	0.148	0.312	0.920
C	39,045	758,000	237	978	4678	43,661	54,100,000
S_1	0.413	0.108	0.026	0.247	0.407	0.599	0.877
S_2	0.103	0.041	0.001	0.046	0.099	0.177	0.393
S_3	0.166	0.089	0.000	0.046	0.154	0.326	0.907
S_4	0.026	0.028	0.000	0.002	0.017	0.078	0.447
S_5	0.292	0.104	0.011	0.131	0.285	0.473	0.958
SY_1	0.058	0.053	0.000	0.006	0.045	0.153	0.879
SY_2	0.458	0.173	0.000	0.169	0.460	0.736	0.946
SY_3	0.187	0.115	0.000	0.046	0.161	0.416	0.931
SY_4	0.281	0.152	0.002	0.074	0.257	0.567	0.960
SY_5	0.016	0.009	0.001	0.008	0.015	0.029	0.275
SY_{123}	0.703	0.152	0.024	0.418	0.726	0.910	0.991
Z_1	9.513	1.192	6.170	7.903	9.373	11.55	18.85

Table 1 continued

Variable	Mean	SD	Percentiles				
			5th	25th	50th	75th	95th
Z_2	0.059	0.065	0.000	0.008	0.040	0.166	0.882
Z_3	0.009	0.015	0.000	0.000	0.005	0.033	0.472
Tot. Assets	1,050,000	21,000,000	5888	25,382	115,000	1,080,000	1,520,000,000

This table shows the average (Mean), standard deviation (SD), the median, and the 5 and 95 percentiles (p5 and p95, respectively). All nominal variables are measured in thousands of 2005 US dollars. The data used for estimation include 60,868 year-bank observations for 7,473 different banks. The output variables are: household and individual loans (Y_1), real estate loans (Y_2), loans to business and other institutions (Y_3), federal funds sold and securities purchased under agreements to resell (Y_4), and other assets (Y_5). The input variables are: labor quantity (X_1), premises and fixed assets (X_2), purchased funds (X_3), interest-bearing transaction accounts (X_4), and non-transaction accounts (X_5). For each input X_j its price, W_j, is computed by dividing total expenses by the corresponding input quantity. Rev is total revenues, and R_q represents the revenue shares for each output category. Likewise, S_k represents cost shares for each input category and SY_q represents output share for each output category. $SY_{123} = (Y_1 + Y_2 + Y_3)/(Y_1 + Y_2 + Y_3 + Y_4 + Y_5)$. Z_1 is log of equity, Z_2 is a proxy for off-balance sheet activities (noninterest income over total income), and Z_3 equals nonperforming loans over total assets

On the cost side, labor input accounts for 41.3 % of total costs (S_1), non-transaction accounts expenditures represent 29.2 % (S_5), premises and fixed assets 10.3 % (S_2), purchased funds 16.6 % (S_3), and transaction accounts 2.6 % (S_4). In addition, Table 1 also reports output shares for each output category, SY_i; and the share of total loans on total output, SY_{123}. In addition, we include the log of total equity capital (Z_1) as a a quasi-fixed netput, a proxy for off-balance sheet activities (Z_2: noninterest income over total income), and a proxy for output quality (Z_3: non performing loans).

4 Empirical results

Both the SPG and the PG models fit the data quite well.[21] The PG model is a restricted version of the SPG model. Then, we can test if the data support the parametric restrictions imposed by the PG model. Using the specification test of Li and Racine (2010), we reject the PG model in favor of the SPG model.[22] For completeness, we present economies of scale and TFP growth components estimates from all three models.

4.1 Estimated functional coefficients

Table 2 presents summary statistics for the estimated functional coefficients. Figure 1 depicts their empirical distributions. The functional coefficients from the SPG model are fully nonparametric and have clear economic meaning: $\beta_0(\cdot)$ captures technical

[21] The R^2 values for these models are 0.978 and 0.916, respectively.The fit of the PL model (R^2) is not comparable to the other two models because the dependent variables are not the same. The R^2 of the PL model is 0.847.

[22] We use 99 bootstrap replications which gave a p-value of almost zero.

Table 2 Estimated functional coefficients

Parameter	Mean	SD	Percentiles				
			5th	25th	50th	75th	95th
Panel A: SPG model ($R^2 = 0.978$)							
β_0	0.001	0.018	−0.022	−0.007	0.000	0.008	0.025
β_1	0.456	0.087	0.343	0.402	0.449	0.504	0.594
β_2	0.044	0.048	−0.004	0.025	0.039	0.057	0.106
β_3	0.164	0.074	0.066	0.118	0.158	0.203	0.288
β_4	0.023	0.021	−0.002	0.015	0.023	0.032	0.050
γ_1	0.059	0.043	−0.003	0.039	0.060	0.080	0.114
γ_2	0.347	0.121	0.163	0.270	0.346	0.418	0.543
γ_3	0.124	0.062	0.041	0.085	0.118	0.155	0.225
γ_4	0.135	0.059	0.052	0.100	0.132	0.166	0.229
γ_5	0.047	0.040	−0.006	0.029	0.046	0.064	0.102
φ_1	0.075	0.086	−0.048	0.033	0.076	0.119	0.193
φ_2	0.202	0.256	−0.106	0.077	0.174	0.296	0.615
φ_3	0.046	1.414	−1.040	−0.226	0.071	0.351	1.155
Panel B: PG model ($R^2 = 0.916$)							
β_0	−0.010	0.007	−0.021	−0.014	−0.009	−0.005	0.000
β_1	0.465	0.069	0.360	0.417	0.461	0.509	0.586
β_2	0.036	0.014	0.015	0.026	0.035	0.044	0.058
β_3	0.176	0.050	0.094	0.144	0.175	0.207	0.257
β_4	0.026	0.010	0.009	0.019	0.026	0.032	0.041
γ_1	0.053	0.015	0.027	0.045	0.054	0.063	0.076
γ_2	0.342	0.123	0.122	0.264	0.352	0.428	0.524
γ_3	0.130	0.046	0.052	0.101	0.131	0.160	0.202
γ_4	0.154	0.054	0.067	0.120	0.153	0.188	0.243
γ_5	0.044	0.011	0.026	0.037	0.044	0.051	0.061
φ_1	0.062	0.031	0.012	0.041	0.061	0.082	0.114
φ_2	0.143	0.126	−0.041	0.054	0.131	0.218	0.371
φ_3	0.057	0.147	−0.178	−0.021	0.064	0.148	0.271
Panel C: PL model ($R^2 = 0.8474$)							
β_0	0.000	0.005	−0.007	−0.003	0.000	0.004	0.009
β_1	0.392	0.094	0.251	0.325	0.384	0.450	0.561
β_2	0.021	0.040	−0.050	−0.003	0.023	0.046	0.082
β_3	0.312	0.091	0.163	0.256	0.313	0.368	0.460
β_4	0.024	0.020	−0.010	0.012	0.025	0.038	0.056
γ_1	0.103	0.047	0.020	0.075	0.107	0.134	0.174
γ_2	0.471	0.153	0.211	0.375	0.476	0.574	0.712
γ_3	0.149	0.076	0.024	0.100	0.150	0.199	0.271
γ_4	0.228	0.106	0.058	0.162	0.227	0.296	0.401
γ_5	0.043	0.029	−0.002	0.024	0.042	0.060	0.089

Table 2 continued

Parameter	Mean	SD	Percentiles				
			5th	25th	50th	75th	95th
φ_1	−0.040	0.035	−0.101	−0.061	−0.038	−0.017	0.012
φ_2	1.227	0.290	0.785	1.057	1.219	1.396	1.709
φ_3	1.430	0.666	0.400	1.005	1.445	1.910	2.415

The data used for estimation include 60,868 year-bank observations for 7,473 different banks from 2001 to 2010. SPG model, PG model, and PL model correspond to the estimation of the functional coefficients in Eq. (5) using the semiparametric smooth coefficient model (SPSCM), a parametric translog growth model, and the parametric translog model in levels, respectively

change (TC) or shifts of the cost function (percentage change in cost, *ceteris paribus*); each $\beta_k(\cdot)$ represents the cost elasticity of input price k; the $\gamma_q(\cdot)$ coefficients are the cost elasticity of outputs and are uniquely related to economies of scale; and the $\varphi_p(\cdot)$ coefficients measure the effects of the control variables on the percentage change in cost.

It can be seen from Table 2 that mean TC is almost zero in SPG and PL models. However looking at the mean might give a misleading impression. A closer look at TC in the PL model shows that it is almost zero for each percentile while for the SPG model there is technical progress (negative TC) for almost half of the banks while for the other half there is technical regress. These opposing forces lead to a technical regress of 0.1 % per annum. On the other hand, the PG model shows technical progress ranging from 2.1 % to 0.00 % for more than 75 % of the banks. On average, technical progress took place at an annual rate of 1 %.

By Shephard's lemma, the coefficients $\hat{\beta}_k(\cdot) = \partial \ln C / \partial \ln W_k$ are cost shares estimates for each input, $k = 1, 2, 3, 4$. The mean and the empirical distributions of $\hat{\beta}_k(\cdot)$ from the SPG and PG models are similar and closely follow the observed cost shares shown in Table 1. Thus, negative values of $\hat{\beta}_k(\cdot)$ $(k = 1, 2, 3, 4)$ indicate violations of regularity conditions since cost shares are positive. Tables 3 and 4 show that most $\hat{\beta}_k(\cdot)$ are positive. For the SPG model the violations are less than 0.5 % for $\hat{\beta}_1(\cdot)$ and $\hat{\beta}_3(\cdot)$, and around 6 % for $\hat{\beta}_2(\cdot)$ and $\hat{\beta}_4(\cdot)$. For the PG model, the violations are less than 0.5 % for all $\hat{\beta}_k(\cdot)$. The PL model has the largest number of violations, 27 % and 11.5 % for $\hat{\beta}_2(\cdot)$ and $\hat{\beta}_4(\cdot)$, respectively. Our results imply that controlling for unobserved heterogeneity using either the SPG or the PG models reduces the number of violations significantly.

Estimated cost elasticity with respect to each output is given by $\hat{\gamma}_q(\cdot)$. We expect these coefficients to have a positive sign since both marginal and average costs are positive. This is the case for most observations. Estimates of $\gamma_q(\cdot)$ from the PG model are negative for less than 0.8 % of all observations. The SPG model shows negative values for $\hat{\gamma}_2(\cdot)$ to $\hat{\gamma}_4(\cdot)$ for less than 1.2 % of the observations. The corresponding values for $\hat{\gamma}_1(\cdot)$ and $\hat{\gamma}_5(\cdot)$ are 5.6 % and 6.25 %. Overall, the PL model gives negative values for $\hat{\gamma}_k(\cdot)$ for less than 6 % of the observations. The small number of violations for the SPG model is remarkable since its estimated coefficients are fully nonparametric functions. Better performance of the PG model in this regard suggests that the PG

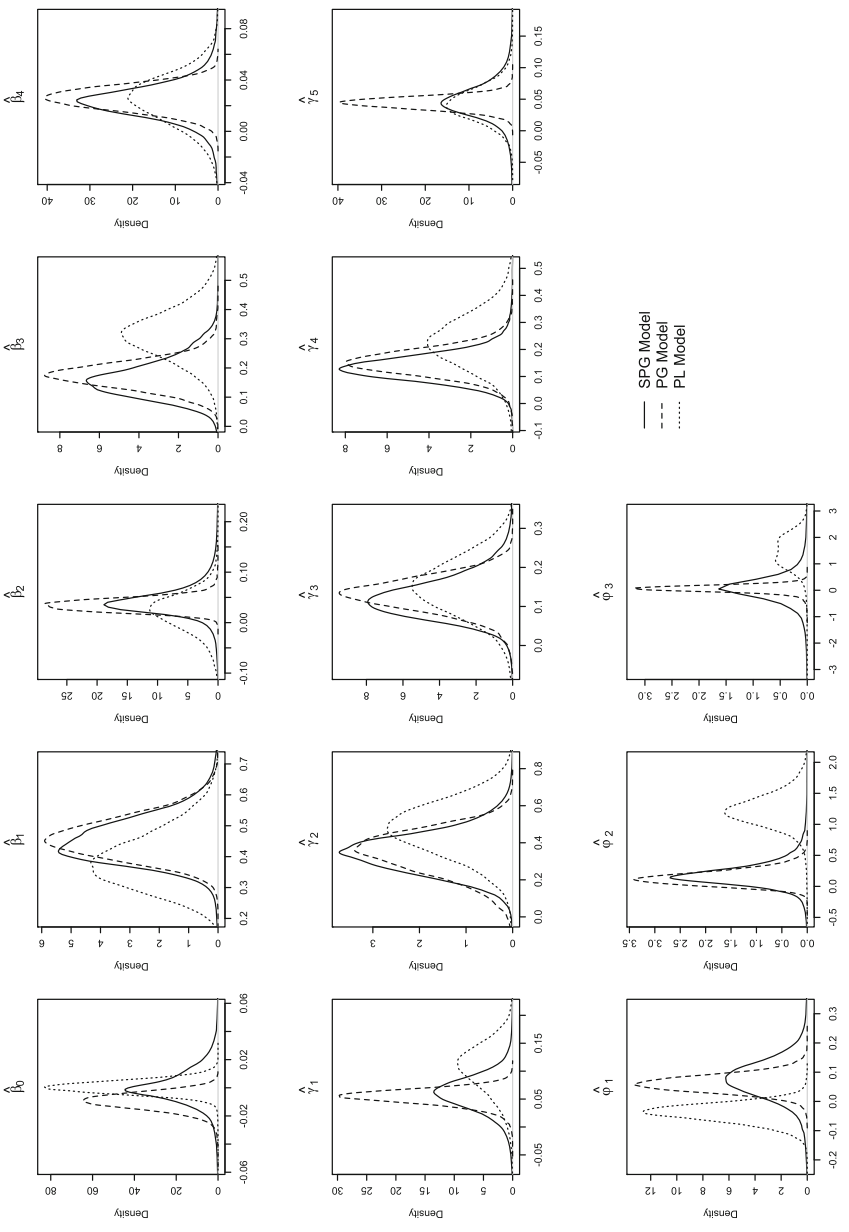

Fig. 1 Density plots of estimated functional coefficients

Table 3 Theoretical monotonicity restrictions

	w_1 (%)	w_2 (%)	w_3 (%)	w_4 (%)	w_5 (%)
% of observations that are non-decreasing in input prices					
SPG model	99.96	94.18	99.47	93.86	99.88
PG model	100.00	99.90	99.97	99.45	100.00
PL model	100.00	72.68	99.82	88.57	98.62
	y_1 (%)	y_2 (%)	y_3 (%)	y_4 (%)	y_5 (%)
% of observations that are non-decreasing in outputs					
SPG model	94.39	99.69	98.87	99.39	93.74
PG model	99.65	99.23	99.21	99.51	99.99
PL model	97.52	99.60	97.09	98.12	93.98

This table shows the percentage of observations for which total cost is non-decreasing in input prices and output quantities. The data used for estimation include 60,868 year-bank observations for 7,473 different banks from 2001 to 2010. SPG model, PG model, and PL model correspond to the estimation of the functional coefficients in Eq. (5) using the semiparametric smooth coefficient model (SPSCM), a parametric translog growth model, and the parametric translog model in levels, respectively

model is a good approximation of the underlying technology. We think that the number of sign violations for the estimated coefficients $\hat{\beta}_k(\cdot)$ and $\hat{\gamma}_k(\cdot)$ in the SPG and PG models are quite low. Thus, we do not restrict their signs in estimating the models.[23]

The $\hat{\varphi}_p(\cdot) = \partial \hat{f}/\partial Z_p$ coefficients capture the effects of log of equity capital (Z_1), off-balance sheet activities (Z_2), and non-performing loans (Z_3) on total cost growth. More specifically, $\hat{\varphi}_1(\cdot)$ represents cost elasticity of equity capital, while $\hat{\varphi}_2(\cdot)$ and $\hat{\varphi}_3(\cdot)$ represent semi cost elasticity of off-balance sheet activities and non-performing loans, respectively. For the SPG and PG models, the estimated mean effects are all positive, indicating that increases in any of these variables to produce a given level of outputs increase total cost, *ceteris paribus*.[24]

It is noteworthy that all the estimated functional coefficients are well behaved as shown by their empirical distributions (Fig. 1). This is particularly remarkable for the nonparametric functional coefficients of the SPG model since they are completely unconstrained. The functional coefficients in all three models are observation-specific. Thus, we compute their standard errors using the wild bootstrap method of Hardle and Mammen (1993). Figure 2 shows the observation-specific functional coefficient estimates from the SPG model along with their 95 % confidence intervals. Unreported results show a similar pattern for the PG and PL models.

[23] Such an exercise is beyond the scope of the present paper. We leave this for future research.

[24] The positive relationship between the growth in variable cost and equity capital suggests that to produce a given output vector with more equity capital banks also spend more on variable inputs. Apart from possible agency problems, two potential and important explanations could be that to do so banks spend more on either reducing or taking on more risk. If nonperforming loans are a valid proxy for realized credit risk, the positive relationship between variable cost and nonperforming loans seems to be consistent with the latter explanation. This discussion is not central to our paper and since it deserves a deeper analysis we left it for future research.

To understand these figures, consider the plot for the estimated functional coefficients $\hat{\beta}_0(\cdot)$ in Fig. 2. We plot $\hat{\beta}_0(\cdot)$ against $\hat{\beta}_0(\cdot)$ such that all the coefficients $\hat{\beta}_0(\cdot)$ lie

Table 4 Hypothesis testings for the coefficient estimates at the 5 % level

	β_0 (%)	β_1 (%)	β_2 (%)	β_3 (%)	β_4 (%)
SPG model					
Significantly positive	24.53	99.93	83.21	98.67	82.54
Insignificant	48.82	0.03	14.20	1.02	14.83
Significantly negative	26.66	0.03	2.60	0.31	2.63
PG model					
Significantly positive	1.04	100.00	98.33	99.88	96.89
Insignificant	12.71	0.00	1.67	0.11	3.09
Significantly negative	86.26	0.00	0.00	0.01	0.02
PL model					
Significantly positive	30.23	100.00	63.79	99.57	79.00
Insignificant	46.20	0.00	16.31	0.37	15.59
Significantly negative	23.57	0.00	19.90	0.07	5.40
	γ_1 (%)	γ_2 (%)	γ_3 (%)	γ_4 (%)	γ_5 (%)
SPG model					
Significantly positive	87.58	99.39	97.77	98.72	82.39
Insignificant	9.57	0.40	1.57	0.90	14.87
Significantly negative	2.85	0.21	0.66	0.38	2.75
PG model					
Significantly positive	97.34	98.83	98.60	99.19	99.21
Insignificant	0.74	0.96	0.57	0.79	11.46
Significantly negative	0.02	0.44	0.44	0.24	0.00
PL model					
Significantly positive	96.49	99.47	95.98	97.72	85.10
Insignificant	1.81	0.25	2.00	0.81	13.21
Significantly negative	1.69	0.28	2.02	1.47	1.69
	ϕ_1 (%)	ϕ_2 (%)	ϕ_3 (%)		
SPG model					
Significantly positive	69.26	63.10	20.94		
Insignificant	25.89	32.98	64.83		
Significantly negative	4.85	3.92	14.23		
PG model					
Significantly positive	88.34	50.01	10.46		
Insignificant	11.46	49.37	87.72		
Significantly negative	0.21	0.62	1.82		
PL model					
Significantly positive	1.37	99.58	94.57		

Table 4 continued

	ϕ_1 (%)	ϕ_2 (%)	ϕ_3 (%)
Insignificant	29.50	0.38	4.74
Significantly negative	69.13	0.04	0.69

This table shows the percentage of observations for which the indicated functional coefficient is significantly positive, insignificant, or significantly negative for each model. The data used for estimation include 60,868 year-bank observations for 7,473 different banks from 2001 to 2010. SPG model, PG model, and PL model correspond to the estimation of the functional coefficients in Eq. (5) using the semiparametric smooth coefficient model (SPSCM), a parametric translog growth model, and the parametric translog model in levels, respectively

along the 45° line. The points above (below) the 45° line represent the upper (lower) bound of each confidence interval. Therefore, for each $\beta_0(\cdot)$ on the 45° line we can see an observation-specific confidence interval. If the horizontal line at zero passes inside of the confidence bounds for any given observation, then $\beta_0(\cdot)$ for this observation is statistically insignificant. Conversely, if the horizontal line at zero passes outside of the confidence bounds, then $\beta_0(\cdot)$ for this observation is statistically significant. In addition, if the lower (upper) bound lies above (below) zero, then the coefficient for this observation is significantly positive (negative). In general, the confidence intervals for each of the functional coefficients are quite tight, although they become wider at the tails.

4.2 Economies of scale

We present summary statistics for economies of scale estimates using the SPG, PG, and PL models in Panels A, B, and C of Table 5. Figures 3 and 4 show their empirical distributions. We find economically and statistically significant scale economies, as measured by RTS, for most banks. For the SPG model (Panel A), mean RTS estimate is 1.42, indicating that if outputs were to increase by 10 %, total costs will increase approximately by 7 % (10 % $\times$ 1/1.42). Estimates of mean RTS from the PG model (Panel B) is 1.40. These results indicate increasing RTS (IRTS) for most banks. The mean RTS from the PL model (Panel C) is 1.01, indicating constant RTS (CRTS). The empirical distribution of estimated RTS from the SPG or PG models show decreasing RTS for less than 2 % of the observations. In contrast, estimated RTS from the PL model show decreasing RTS (DRTS) for about 42 % of the observations.

It can be seen from Table 5 that RTS estimates decrease with bank size.[25] Mean RTS for small, medium and big banks are 1.56, 1.34, and 1.18 from the SPG model and 1.54, 1.32, and 1.10 from the PG model. Mean RTS for the top 100 banks is also lower than for all other banks. Estimated RTS from the SPG and PG models show that almost all medium and small banks exhibit IRTS. More than 75 % of big banks and

[25] We classify banks by total assets as follows: Big (assets > $1 billion), medium ($100 million < assets < $1 billion), and small (assets < $100 million) banks. The top 100 banks correspond to the 100 biggest banks by assets each year. This classification is typically used in the literature (see Feng and Serletis 2009).

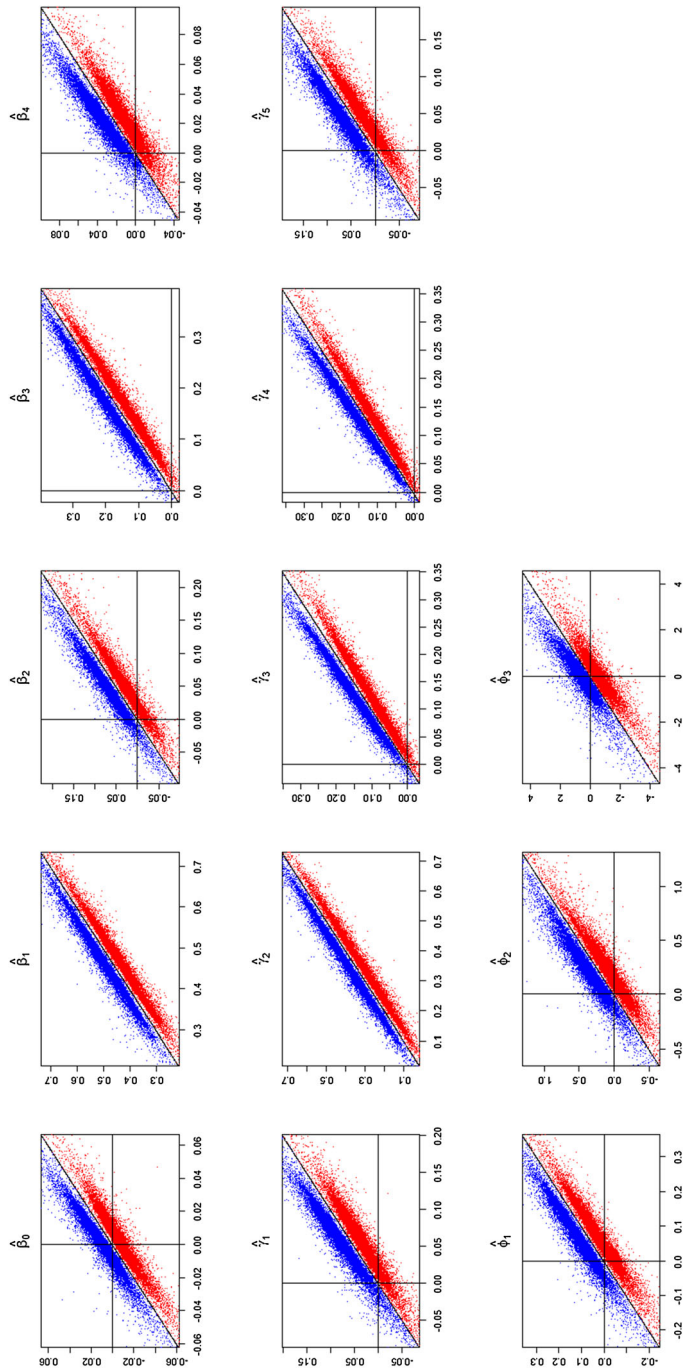

Fig. 2 Confidence intervals for bootstrapped standard errors of SPG model

Table 5 Summary statistics for returns to scale estimates by bank size

Bank size	Mean	SD	Percentiles				
			5th	25th	50th	75th	95th
Panel A: SPG model							
TOP 100	1.127	0.276	0.799	0.959	1.086	1.237	1.577
Big	1.180	0.234	0.903	1.044	1.144	1.266	1.581
Medium	1.335	0.164	1.130	1.239	1.311	1.399	1.626
Small	1.561	0.263	1.263	1.389	1.512	1.670	2.020
All	1.425	0.247	1.135	1.274	1.375	1.532	1.860
Panel B: PG model							
TOP 100	1.011	0.087	0.859	0.957	1.019	1.065	1.144
Big	1.103	0.095	0.920	1.052	1.112	1.162	1.240
Medium	1.323	0.098	1.177	1.255	1.317	1.381	1.491
Small	1.541	0.141	1.356	1.436	1.516	1.621	1.810
All	1.403	0.174	1.152	1.287	1.386	1.502	1.722
Panel C: PL model							
TOP 100	1.036	0.050	0.964	1.002	1.031	1.064	1.132
Big	1.022	0.043	0.963	0.993	1.016	1.043	1.100
Medium	1.008	0.034	0.953	0.986	1.007	1.028	1.064
Small	1.009	0.035	0.953	0.986	1.009	1.032	1.067
All	1.009	0.035	0.954	0.987	1.008	1.031	1.067

This table shows summary statistic for RTS estimates. The data used for estimation include 60,868 year-bank observations for 7,473 different banks from 2001 to 2010. SPG model, PG model, and PL model correspond to RTS estimates computed after estimating Eq. (5) using the semiparametric smooth coefficient model (SPSCM), a parametric translog growth model, and the parametric translog model in levels, respectively. TOP 100 corresponds to RTS estimates for the 100 biggest banks. Banks size categories are: Big (assets >$1 billion), medium ($100 million < assets < $1 billion), and small (assets <$100 million) banks

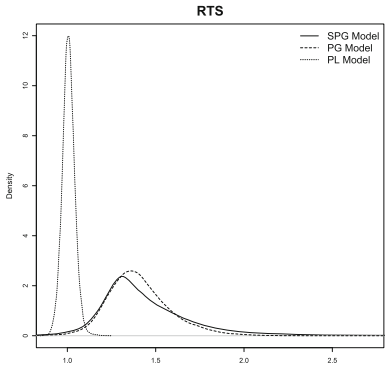

Fig. 3 Density plots for RTS estimates for all banks

more than 50 % of the top 100 banks also exhibit IRTS. Estimated RTS from the PL model show IRTS for about 50 % of banks.[26]

[26] RTS estimates from the SPG model has more extreme values than those from the PG model. This is expected from a nonparametric model. That is, the distribution of RTS from the PG model is quite tight.

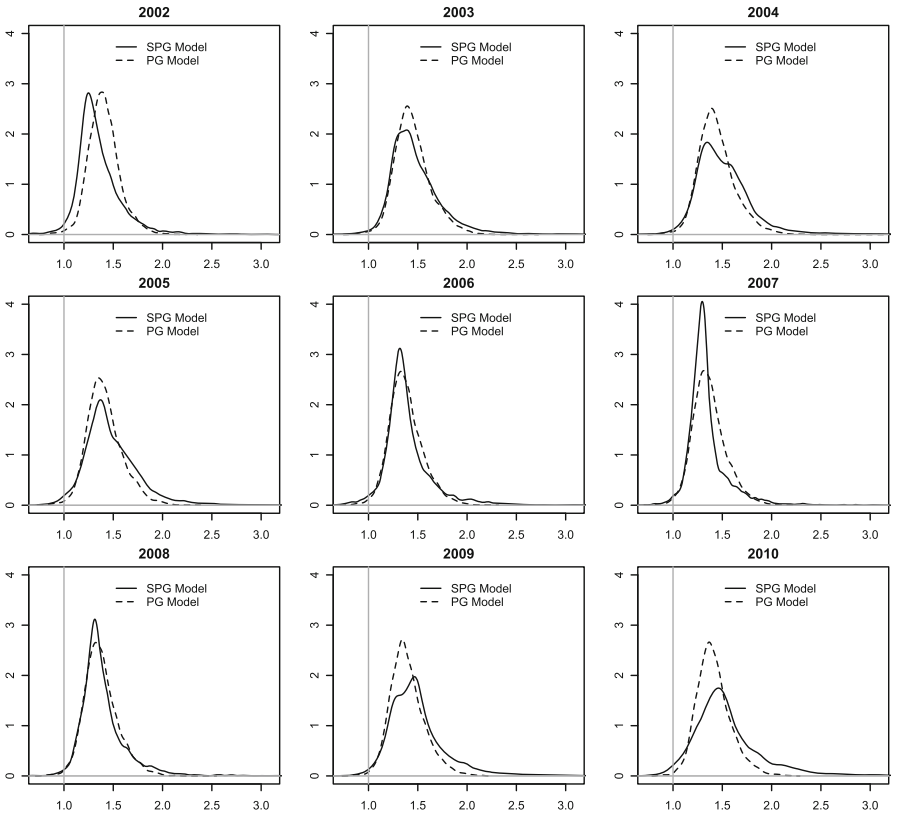

Fig. 4 RTS estimates over time

The estimates of RTS from the SPG and the PG model are directly comparable in the sense that both models control for fixed bank-specific effects. This is, however, not the case with the PL model. Thus, differences between estimated RTS from the PL model and those from the other two models might be attributed to the omission of bank-specific effects in the specification of the PL model. Estimated RTS from the PL model are statistically less significant and economically smaller than those from the PG model, thereby indicating that failure to account for time-invariant unobserved bank heterogeneity conceals the evidence of substantial scale economies for most banks.

Now we investigate the statistical significance of our RTS estimates. In Fig. 5, we report estimates of the 95 % confidence intervals for RTS by bank size for each model. In Table 6, we report the percentage of observations exhibiting statistically significant increasing, decreasing, or constant RTS. For each plot in Fig. 5, points above (below) the 45° line represent the upper (lower) bounds of observation-specific confidence intervals. Points on the 45° line are the estimated RTS. If the horizontal line at unity passes inside of the confidence bounds for any given observation, then RTS estimate for this observation equals unity (CRTS) statistically. If the lower (upper) bound lies

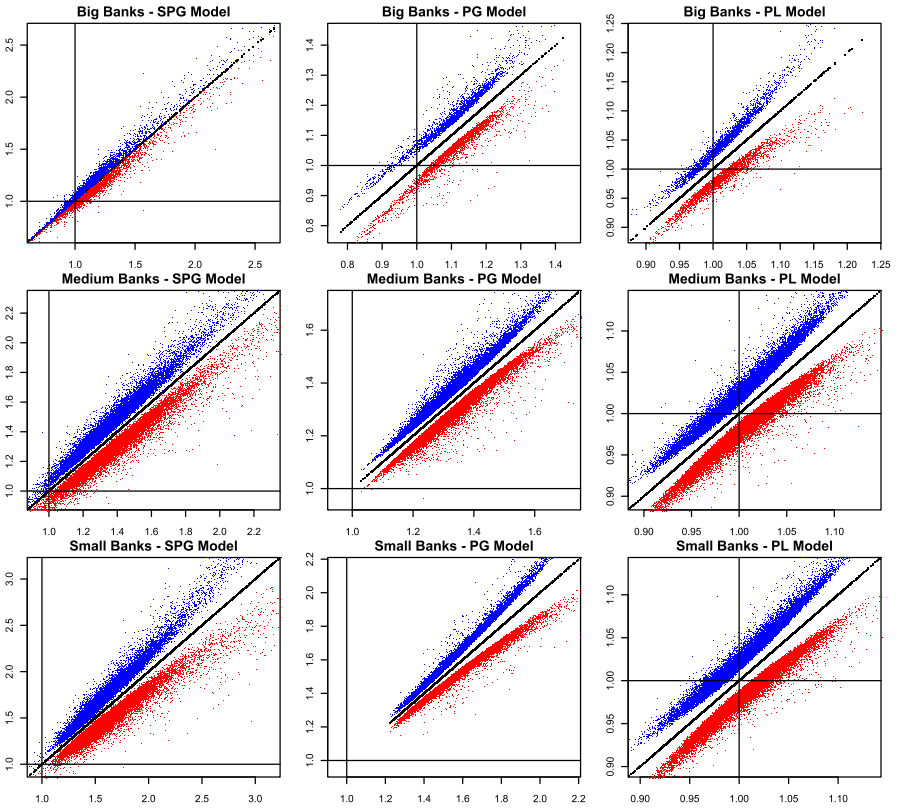

Fig. 5 Confidence intervals for RTS estimates for each model

above (below) unity, then the RTS estimate for this observation is significantly greater (less) than unity, indicating IRTS (DRTS).

The confidence intervals in Fig. 5 are consistent with the existence of IRTS for most banks reported above. In the SPG model, 98 % of the observations for medium and small banks show presence of IRTS. The corresponding values for big and the top one hundred banks are 75 % and 65 %, respectively (see Table 6). Only 2 % of the observations for medium and small banks show presence of CRTS and none shows evidence of DRTS. Only 13 % of the big banks and 26 % of the top one hundred banks show evidence of DRTS.

The evidence from the PG model is consistent with the results from the SPG model, except for the top one hundred banks. The estimates from the PG model show evidence of IRTS for 100 %, 74 %, and 23 % of medium and small banks, big banks, and the top one hundred banks, respectively. In contrast, estimates from the PL model show no evidence of IRTS for most banks.

We plot the relation between estimated RTS and bank size in Fig. 6 for the SPG and PG models. Estimated RTS from the SPG and PG models decrease monotonically as bank size increases. There is a clear inverse relation between RTS and bank size—measured by size deciles. The interquartile range of estimated RTS tend to be wider

Table 6 Percentage of big banks with IRTS, DRTS, or CRTS

Year	N	SPG model			PG model			PL model		
		IRTS (%)	DRTS (%)	CRTS (%)	IRTS (%)	DRTS (%)	CRTS (%)	IRTS (%)	DRTS (%)	CRTS (%)
Medium and small banks										
2002	5021	95	0	4	100	0	0	14	32	54
2003	6181	99	0	1	100	0	0	29	19	52
2004	6175	99	0	1	100	0	0	35	15	50
2005	6066	99	0	1	100	0	0	28	21	51
2006	5852	98	0	1	100	0	0	19	30	51
2007	5554	99	0	1	100	0	0	15	35	50
2008	5161	99	0	1	100	0	0	29	20	51
2009	4718	99	0	1	100	0	0	48	7	45
2010	4364	98	0	2	100	0	0	65	3	33
Avg	5455	98	0	2	100	0	0	31	20	49
Big banks										
2002	260	68	23	8	70	6	24	23	9	68
2003	309	83	12	5	75	5	20	41	5	54
2004	340	85	7	8	76	5	19	39	8	54
2005	355	70	12	18	73	6	21	27	17	55
2006	364	59	18	23	69	7	24	14	31	55
2007	349	65	13	22	68	6	26	10	39	51
2008	323	80	11	9	72	7	21	24	21	54
2009	306	79	10	10	76	5	19	49	8	43
2010	268	72	14	14	82	3	15	65	3	32
Avg	319	73	13	13	74	5	21	32	16	52

Table 6 continued

Year	N	SPG model			PG model			PL model		
		IRTS (%)	DRTS (%)	CRTS (%)	IRTS (%)	DRTS (%)	CRTS (%)	IRTS (%)	DRTS (%)	CRTS (%)
Top 100 banks										
2002	100	39	51	10	25	15	60	30	1	69
2003	100	71	25	4	26	15	59	55	1	44
2004	100	84	15	1	22	17	61	49	3	48
2005	100	73	18	9	12	22	66	30	9	61
2006	100	60	30	10	11	25	64	22	21	57
2007	100	51	29	20	14	22	64	15	29	56
2008	100	71	24	5	18	21	61	30	17	53
2009	100	70	21	9	29	14	57	52	6	42
2010	100	70	17	13	52	9	39	65	2	33
Avg.	100	65	26	9	23	18	59	39	10	51

This table shows the percentage of observations with statistically significant increasing (IRTS), decreasing (DRTS), and constant (CRTS) RTS. A specific RTS estimate indicates IRTS, DRTS, or CRTS if its 95 % confidence interval lies entirely above, entirely below, or includes one, respectively. The data used for estimation include 60,868 year-bank observations for 7,473 different banks from 2001 to 2010. SPG model, PG model, and PL model correspond to RTS estimates computed after estimating Eq. (5) using the semiparametric smooth coefficient model (SPSCM), a parametric translog growth model, and the parametric translog model in levels, respectively. TOP 100 corresponds to RTS estimates for the 100 biggest banks. Banks size categories are: Big (assets >$1 billion), medium ($100 million < assets < $1 billion), and small (assets <$100 million) banks

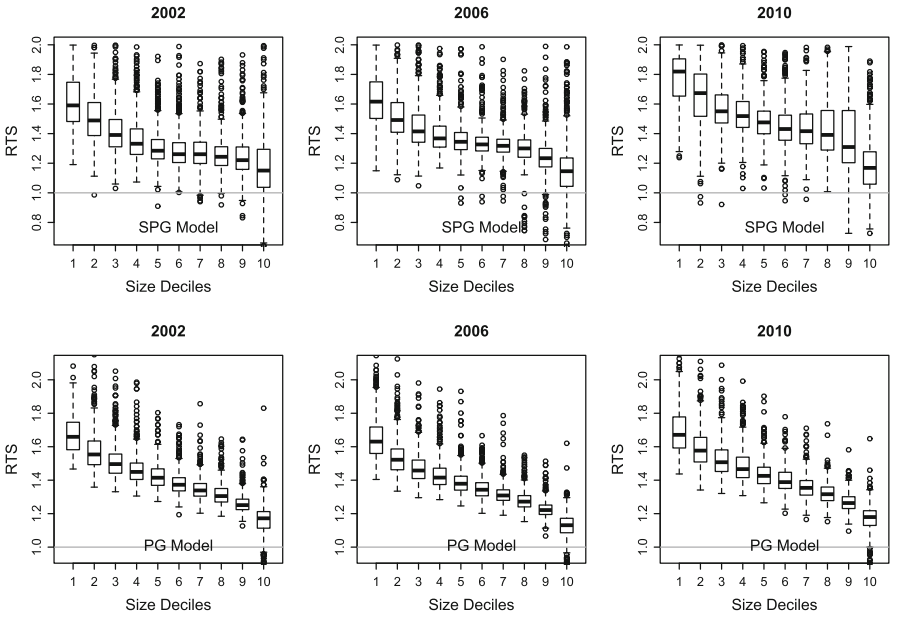

Fig. 6 RTS estimates by bank size deciles based on total assets

for the smallest and the biggest banks. Unreported results show that the empirical distribution of bank size is almost the same for the biggest one hundred banks that exhibit either increasing or decreasing RTS. Thus, for the biggest one hundred banks (as measured by total assets) estimated RTS are unrelated to bank size, suggesting that some of these banks will continue growing irrespective of their actual size.

Table 7 presents additional results concerning the distribution of RTS for the biggest ten banks in our sample from 2001 to 2010. RTS estimates for these banks are lower for the crisis years (2007–2009). We investigate this issue further by estimating the SPG model using dummy variables for the crisis years and including in the sample only banks with more than $500 million in assets (results not shown). We find that estimated RTS are also lower during the crisis years but the reduction in their magnitude is not as pronounced. Thus, there seem not to be a significant change in banks' cost structure during those years.

Based on transition probability matrices (not reported), we find that estimated RTS from the SPG model are stable over time. The likelihood of remaining within the same RTS decile or switching to the two adjacent RTS deciles are around 70 % for the SPG model and 90 % for the PG model, suggesting that the RTS estimates from the PG model are more stable over time. This result may stem from the parametric assumptions underlying the PG model. In contrast, estimated RTS in the SPG model are fully nonparametric. Nonetheless, the probabilities of switching across deciles that are farther away from each other are small for both models, indicating that there are no large swings on the estimated RTS. Unreported piece-wise Spearman rank correlations show that the rank correlation for adjacent years are high for both the SPG and the PG model (above 0.75 and 0.98, respectively). These results are consistent with the

Table 7 Summary statistics for nonparametric RTS estimates for the ten biggest banks

Year	RTS	Mean	SD	Percentiles				
				5th	25th	50th	75th	95th
2002	Lower bound	1.385	0.418	0.860	1.053	1.308	1.766	1.969
	Estimate	1.393	0.423	0.864	1.055	1.310	1.785	1.977
	Upper bound	1.400	0.427	0.867	1.057	1.312	1.804	1.985
2003	Lower bound	1.441	0.404	0.675	1.234	1.447	1.586	2.046
	Estimate	1.459	0.423	0.676	1.237	1.448	1.592	2.098
	Upper bound	1.477	0.442	0.676	1.240	1.450	1.598	2.151
2004	Lower bound	1.376	0.306	1.033	1.190	1.275	1.534	1.919
	Estimate	1.397	0.332	1.036	1.197	1.277	1.536	2.004
	Upper bound	1.418	0.359	1.039	1.200	1.279	1.538	2.125
2005	Lower bound	1.634	0.520	0.927	1.307	1.613	1.808	2.634
	Estimate	1.657	0.525	0.928	1.321	1.631	1.802	2.661
	Upper bound	1.681	0.533	0.930	1.335	1.760	1.832	2.689
2006	Lower bound	1.228	0.501	0.712	0.905	0.998	1.484	2.271
	Estimate	1.242	0.514	0.714	0.910	1.011	1.497	2.301
	Upper bound	1.256	0.527	0.715	0.915	1.023	1.511	2.331
2007	Lower bound	0.758	0.200	0.506	0.651	0.759	0.807	1.217
	Estimate	0.774	0.195	0.524	0.652	0.766	0.808	1.220
	Upper bound	0.790	0.193	0.543	0.652	0.808	0.833	1.224
2008	Lower bound	0.766	0.236	0.429	0.520	0.804	0.925	1.165
	Estimate	0.781	0.228	0.458	0.550	0.804	0.955	1.165
	Upper bound	0.796	0.222	0.487	0.580	0.804	0.985	1.165
2009	Lower bound	1.085	0.246	0.786	0.894	1.084	1.199	1.537
	Estimate	1.093	0.254	0.786	0.901	1.088	1.206	1.568
	Upper bound	1.101	0.263	0.786	0.907	1.092	1.213	1.599
2010	Lower bound	1.153	0.240	0.814	1.055	1.083	1.257	1.608
	Estimate	1.157	0.242	0.817	1.058	1.087	1.258	1.619
	Upper bound	1.162	0.245	0.819	1.062	1.091	1.258	1.629
Total	Lower bound	1.199	0.445	0.597	0.860	1.141	1.484	1.969
	Estimate	1.214	0.453	0.614	0.864	1.143	1.497	2.004
	Upper bound	1.228	0.462	0.632	0.867	1.145	1.510	2.094

This table shows summary statistic for nonparametric RTS estimates for the ten biggest banks in the sample each year obtained using the SPG Model. The data used for estimation include 60,868 year-bank observations for 7,473 different banks from 2001 to 2010. SPG model estimates correspond to RTS estimates computed after estimating Eq. (5) using the semiparametric smooth coefficient model (SPSCM). Upper and lower bounds correspond to the bounds of 95 % confidence interval around RTS estimates. Among the top ten biggest banks each year are State Street Bank and Trust Company, CitiBank, US Bank, Wachovia, HSBC, Wells Fargo, Bank of America, Bank of New York Mellon, Fleet national Bank, Suntrust Bank, Keybank, PNC Bank, Regions Bank, JP Morgan Chase, and Citizens Bank

transition probabilities matrix analysis discussed before and may stem from the fact that the parametric specification imposes stronger restrictions on the dynamics of RTS in contrast to the fully nonparametric RTS from the SPG model.

4.3 TFP growth and its components

Unlike previous studies, our framework allows us to tie the functional coefficients of our models to TFP growth components, allowing us to investigate the main forces explaining US banking industry growth. As we show below, our results show that increasing returns to scale played an important role in explaining TFP growth of the US banking industry. By comparison, the role of technical change is smaller. These results are consistent with recent evidence presented by Diewert and Fox (2008) for the US manufacturing industry.

We compute TFP growth (the Divisia) from $T\dot{F}P \equiv \sum_{q=1}^{Q} R_q \dot{Y}_q - \sum_{k=1}^{K} S_k \dot{X}_k$. Note that the Divisia can be computed directly from the data without estimating any econometric model. In order to decompose it into technical change, scale, allocative, control, and random components using Eq. (8), we need to estimate the components econometrically. That is, in an econometric model one estimates the components and then adds them to compute TFP growth. Table 8 summarizes the results and Fig. 7 presents density estimates for TFP growth and its components. The first row of each panel in Table 8 corresponds to the estimated TFP annual growth rate. Since the SPG and PG models use a growth formulation, TFP growth rates differ from the Divisia TFP growth rate only by the random component which has a zero mean. This is why TFP growth obtained from adding the components equals the Divisia index. This is, however, not the case in the PL model in which the error term is not the same as the last component of TFP growth in Eq. (8).[27]

Average annual TFP growth rate is 3.1 % with a standard deviation of 8.3 %. Average TFP growth using either the SPG or the PG model equals the average Divisia TFP growth rate but its standard deviation is lower, reflecting the absence of the random component. Average TFP annual growth rate using the PL model is 1.1 % with a standard deviation of 8.4 %. Since the the PL model uses data in logs, it fits total cost changes less accurately, resulting in a large unexplained component (which is zero on average in both the SPG and PG models). Thus, TFP growth rate estimates from the PL model is likely to differ from the Divisia TFP growth rates.[28]

The top left panel of Fig. 8 shows estimated average annual TFP growth rates over time for the Divisia and the three econometric models. TFP growth rates are positive over the sample period for the Divisia, the SPG and PG models. TFP growth rates are higher from 2007 to 2010 than from 2002 to 2006. This pattern coincides with the US financial crisis period. However, except for an increase in the elasticity of cost with respect to real estate loans, there is no apparent change in the cost structure for most banks.[29] This result holds for the Divisia, the SPG and PG models. In contrast, TFP growth rate estimates using the PL model overestimate TFP growth rates from 2002 to 2006 and underestimate TFP growth from 2007 onwards. Compared with 2001

[27] See Kumbhakar and Sun (2012) for a discussion on this issue based on an input distance function formulation.

[28] This finding is not new in the TFP growth literature. For example, see Kumbhakar and Lozano-Vivas (2005) and the references cited in there.

[29] One of the anonymous referees indicated that these results could be of independent interest. We leave a throughout analysis of them for future research.

Table 8 TFP components

Variable	Mean	SD	Percentiles				
			5th	25th	50th	75th	95th
TFP divisia	0.031	0.083	−0.077	−0.009	0.026	0.064	0.150
Panel A: SPG model							
TFP growth	0.031	0.076	−0.057	−0.005	0.024	0.058	0.140
Scale	0.022	0.070	−0.053	−0.009	0.013	0.041	0.120
TC	−0.001	0.018	−0.025	−0.008	0.000	0.007	0.022
Allocative	0.017	0.042	−0.040	−0.005	0.013	0.037	0.087
Exogenous	−0.007	0.025	−0.036	−0.011	−0.004	0.000	0.014
Panel B: PG model							
TFP growth	0.031	0.071	−0.058	−0.005	0.024	0.059	0.139
Scale	0.018	0.063	−0.052	−0.009	0.012	0.037	0.108
TC	0.010	0.007	0.000	0.005	0.009	0.014	0.021
Allocative	0.007	0.038	−0.049	−0.013	0.006	0.026	0.068
Exogenous	−0.004	0.012	−0.020	−0.007	−0.003	−0.000	0.007
Panel C: PL model							
TFP growth	0.011	0.084	−0.106	−0.030	0.008	0.050	0.135
Scale	0.002	0.051	−0.066	−0.017	0.003	0.022	0.065
TC	−0.000	0.005	−0.009	−0.004	−0.000	0.003	0.007
Allocative	0.008	0.057	−0.081	−0.021	0.008	0.038	0.097
Exogenous	0.002	0.043	−0.046	−0.013	−0.000	0.012	0.052

Results from 60,868 bank-year observations for 7,473 different banks for SPG, PG, and PL models

levels, the TFP indexes from the Divisia, the SPG and PG models increase about 33 % by 2010 (see the top right plot in Fig. 8). The PL model's TFP index shows that the increase was only about 11 %. SPG model's average annual TFP growth rate for big, medium, and small banks are 3.9 %, 3.1 %, and 3 %, respectively. For the PG model, the corresponding values are 3.3 %, 3.3 %, and 2.9 %; and for the PL model they are 1.6 %, 1 %, and 1.1 %.

Taken together, these results suggest that estimating the underlying parameters of the cost function without controlling for unobserved heterogeneity may lead to biased estimates of TFP growth components. In particular, in the PL model the mean difference between the annual total cost change and its estimate is −2 %. This is exactly the difference between the average TFP growth estimated from the PL model and the Divisia. Therefore, without controlling for time-invariant unobserved heterogeneity, the parameter estimates of the underlying cost function are biased. As a consequence, the PL model attributes most of the TFP growth to the random component. Now we focus on the sources of TFP growth by examining each component separately.

4.3.1 Scale economies

The scale component of TFP growth in Eq. (8), $\sum_{q=1}^{Q}(R_q - \gamma_q)\dot{Y}_q$, can be rewritten as $(RTS - 1)\sum_{q=1}^{Q}\gamma_q(\cdot)\dot{Y}_q + \sum_{q=1}^{Q}(R_q - \gamma_q(\cdot)/\Gamma(\cdot))\dot{Y}_q$ where $\Gamma(\cdot) = \sum_q \gamma_q(\cdot)$.

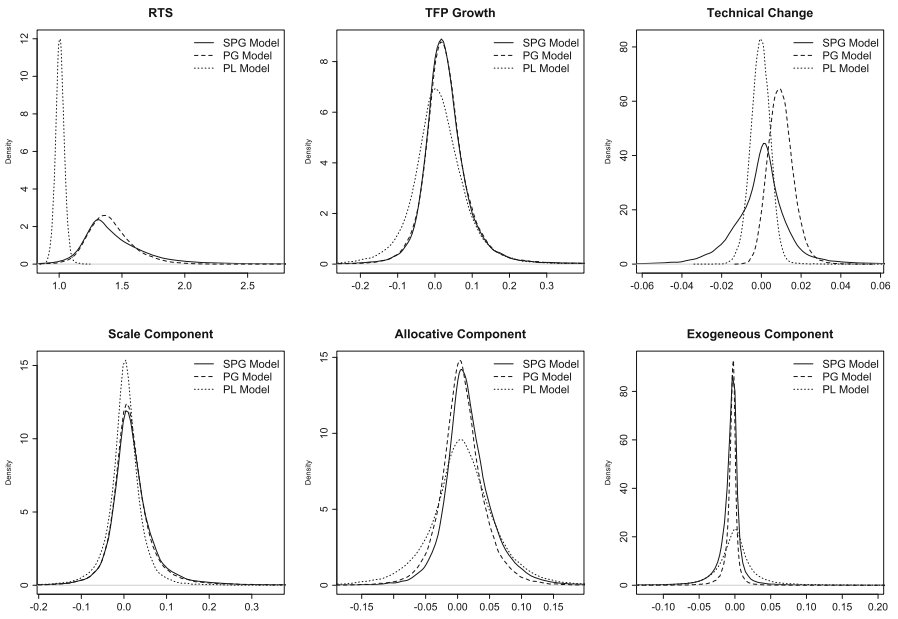

Fig. 7 Density plots for RTS and TFP components

The presence of scale economies (IRTS) contributes positively to TFP growth if output growth rates ($\dot{Y}_q$) weighted by the elasticity of cost with respect to each output (γ_q) are positive. If the marginal cost for each output equals the corresponding output price, then $R_q = \gamma_q / \Gamma_q$ which means $\sum_{q=1}^{Q} (R_q - \gamma_q / \Gamma_q) \dot{Y}_q = 0$. If not, this component will be non-zero and can be interpreted as the mark-up component. The means of this mark-up component for the SPG, PG, and PL models are 0.6 %, 0.3 %, and 0.2 %, respectively.

The mean of the scale component estimates equals 2.2 %, 1.8 %, and 0.2 % for the SPG, PG, and PL models, respectively. Thus, on average, economies of scale contribute positively to TFP growth. Compared with medium and small banks, the contribution of scale economies to TFP growth is higher for big banks. The positive contribution indicates that, on average, $R_q - \gamma_q(\cdot) \geq 0$, which means that the effects of output price changes on total revenue are higher than the corresponding effects of output quantity changes on cost (i.e., a 1 % change in a given output causes a proportional change in total cost that is lower than the proportional change in total revenue caused by a 1 % increase in the corresponding output price.)

The lower-left plot of Fig. 8 shows that the scale components for the SPG and PG models are similar. The SPG scale component tends to grow faster, however, reflecting the higher RTS obtained using the SPG model. This plot also shows that the differences in estimated RTS between the SPG and PG model are small and their contributions to TFP growth are comparable.

4.3.2 Technical change

From Eq. (8), the contribution of TC to TFP growth is $-\beta_0(\cdot) = -\partial f / \partial t$. Estimates of TC from the SPG and the PG models show an annual rate of TC of about -0.1

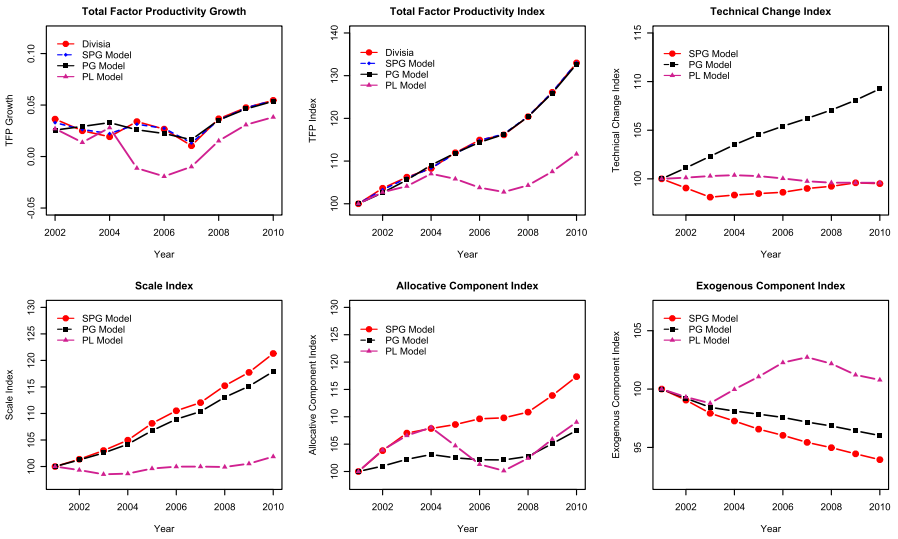

Fig. 8 TFP growth and TFP indexes over time

and 1 %, respectively (see Table 2). Results from these two models also show that big banks experience higher rates of TC than medium and small banks. Mean TC from the SPG model are 0.6 %, −0.1 %, and −0.1 %, for big, medium, and small banks, respectively. The corresponding values from the PG model are 1.7 %, 0.9 %, and 1 %.

These differences lead to different temporal paths of TC as shown in the left-middle plot of Fig. 8. Estimates of TC from the PG model show that banks experience substantial gains in TFP growth. The SPG model suggests that total cost remains essentially unchanged from 2001 to 2010, *ceteris paribus*. Given that the SPG model imposes fewer constraints on the model and therefore on the dynamics of TC, we take the results from this model more seriously.

4.3.3 Allocative component

TFP growth is also affected by deviations of actual input cost shares, S_k, from the optimal input cost shares, β_k. This happens when banks fail to allocate inputs optimally. The allocative component, $AL = \sum_{k=1}^{K-1}(S_k - \beta_k)\dot{\tilde{W}}_k$, captures the contribution of such input misallocations (over- or under-use of inputs) on TFP growth. If $S_k = \beta_k$, $k = 1, \ldots, 4$, then there is no input mis-allocation and therefore $AL = 0$. The sign on AL depends on the extent of input mis-allocation as well as rate of change in relative input prices.

The average contribution of AL in the SPG and PG models are 1.7 % and 0.7 %, with standard deviation of 4.2 % and 3.8 %. For the PL model it is 0.8 % with a standard deviation of 5.7 %. Fig. 8 shows the temporal behavior of AL in the three models. The AL component from the SPG model shows an increase of about 17 % during 2001 to 2010. The corresponding values for the PG and PL models are about 6 % and 7 %, respectively. The AL component from the SPG model show a steady

increase during the sample period. On the other hand, the PG and PL models show increases up to 2004, decreases for 2005 and 2006, and steady gains thereafter. The estimates from all three models indicate allocative efficiency gains from 2007 to 2010.

Gains from AL predicted by the SPG model are 2 % for big banks, 1.6 % for medium and 1.8 % for small banks. The corresponding values from the PG model are 0.8 %, 0.7 %, and 0.7 %, respectively. In contrast, the estimates from the PL model are about 0.8 % regardless of bank size.

4.3.4 Component attributed to environmental factors

TFP growth can also be affected by factors other than TC, allocative efficiency, and scale economies. We consider three such environmental factors, viz., (log) equity capital (Z_1), non-traditional activities (off-balance sheet activities: Z_2), and non-performing loans (Z_3). The contribution of these control/environmental variables to TFP growth is capture by $EX = -\sum_{p=1}^{P} \varphi_p(\cdot)\nabla_t Z_p$ in Eq. (8). The EX values for the SPG and PG models follow a similar pattern over time. Average values of EX from the SPG and PG models are small but negative, -0.7 % and -0.4 % with standard deviations of 2.5 % and 1.2 %, respectively. Thus, these environmental variables contribute negatively to TFP growth. However, their effects are quite small compared with the other components. The last plot in Fig. 8 shows that the cumulative effect of the environmental variables for the 2001–2010 period based on the SPG and PG models is about -6 % and -3 %, respectively. On the other hand, the PL model shows that the EX component contributed positively to TFP growth.

5 Robustness checks

We conduct several robustness checks to examine stability of our results. As pointed out by DeYoung (2010), evidence of scale economies for big banks may be driven by inclusion of very small banks in the estimation process. Thus, we re-estimated the SPG model using data for commercial banks with assets in excess of $500 million. In this case, estimated median scale economies dropped to 1.16 from 1.37 (see Table 5). For banks with assets in the 90th percentile, median scale economies are slightly higher, 1.18. Thus, despite the drop in median scale economies for this subsample, there is still strong evidence of scale economies for even the biggest banks.

We also investigate robustness of our results when the financial crisis of 2007-2009 is controlled. We re-estimated the SPG model including data for commercial banks with assets in excess of $500 million and included a dummy variable for the crisis years. Except for an increase in the elasticity of total costs with respect to real estate loans for the biggest banks, our results indicate no change in the cost structure of banks during the crisis years. RTS estimates in this case are slightly lower but the results are still consistent with substantial scale economies.

Selecting/including output categories appropriately is an important factor in measuring scale economies. Off-Balance-Sheet (OBS) activities account for an important part of non-interest income for banks which suggests treating it as an output and not as a control variable. We conducted a separate estimation for large commercial

banks with assets greater than $500 million including OBS activities (with the level of non-interest income as a proxy) as an additional output.[30] Compared with the results presented previously in Table 5, OBS has a very small impact on our estimates of RTS. Without OBS as an additional output, mean of estimated RTS for banks with total assets greater than $1 billion is 1.18. With OBS as an additional output, the estimate of mean RTS is 1.19. In addition, about 80 % of the estimated RTS are greater than 1.05, indicating that a large proportion of the biggest commercial banks enjoy significant economies of scale. In this sense, the qualitative results of the paper do not change.

In our previous specifications, we treat nonperforming loans as a control variable for risk and measure it as a ratio. However, nonperforming loans could be considered as a quasi-fixed input since to produce more output, especially loans, banks could opt to take on more credit risk. The derivative of cost with respect to the level of nonperforming loans would thus indicate its relationship to cost and give us an indication of how variable cost varies with realized credit risk. For this, we re-estimated the model using nonperforming loans in level as a quasi-fixed input for banks with assets in excess of $500 million. We find that the overall relation between realized credit risk and variable costs is positive, indicating that as banks take on more credit risk, variable costs also increase. The inclusion of nonperforming loans in levels does not change our estimates of RTS significantly. In addition, we find no specific pattern on the relationship between realized credit risk and total cost as a function of bank size.

To investigate the consistency of our scale economies estimates with those recently reported in the literature, we compare them with those in Hughes and Mester (2013) and Wheelock and Wilson (2012). From their preferred model specification, Hughes and Mester (2013) report mean scale economies as high as 1.43 for the biggest US Bank Holding Companies (BHCs) operating in 2010 (see page 573). In comparison, our mean scale economies estimate for the top one-hundred banks is 1.12 and for the ten biggest banks is 1.21 (see Table 7). In addition, Wheelock and Wilson (2012) find that for the 20 biggest BHCs in their sample, scale economies were about 1.25.[31] Thus, our results seems lower than those recently reported in the literature.

We use annual US Call Report data for individual commercial banks. However, several banks operate under a BHC. Hence, a BHC may conduct its subsidiary banks as part of an overall business strategy. Thus, investigating scale economies at the commercial bank level may not capture the overall BHC strategy and cannot account for the dependence between multiple institutions owned by the same BHC. Thus, our empirical results are to be interpreted in the context of commercial banks only. We

[30] We thank two anonymous referees for this suggestion.

[31] Wheelock and Wilson (2012) do not report specific measures of scale economies. However, from their discussion on page 192 we could back up the value of 1.25 for the biggest 20 BHCs. They explain that by reducing the size of a BHC like Citibank by a factor of 0.5012 leads to a decrease in costs by a factor $1 - 0.5952$. This is equivalent to a decrease of $0.4008 = 0.5012 \times 1/1.2504 = 1 - 0.5952$. Equivalently, increasing the size of Citibank by 1 %, would increase its costs by only 0.7997 % ($1\ \% \times 1/1.2504$). Given that the size of Citibank in 2006 (measured by assets) was $1.885 trillion dollars, a 1 % increase in total assets ($18.85 billion) would lead to an increase in total costs by $15.07 billion ($18.85 \times 1/1.2504$) which implies scale economies of $3.77 billion.

re-estimated the model using data for BHCs from 2001 to 2010. Contrary to our results for individual commercial banks, we find that most BHCs exhibit decreasing returns to scale (the results are available from the authors upon request). Median value of RTS for this sample is about 0.726. Notably, for the biggest BHCs (those belonging to the 90th decile by asset size) the RTS is about 0.77. Hence, there are substantial differences between scale economies estimated for individual banks and those for BHCs.

6 Regulatory implications

Despite the wide range of problems recently addressed by enacted financial regulations in the US, policymakers, regulators, academics, and financial market participants are still pondering the idea of capping the size of banks, bringing the issue of existence of scale economies to the fore front of the policy debate. If big banks enjoy substantial scale economies, breaking up the biggest banks or capping their size may impose efficiency losses to the society.

Wheelock and Wilson (2012) estimate that the cost of breaking the four largest US BHCs in existence in 2010 would hover around $79 billion annually. On the other hand, Boyd and Heitz (2012) estimates that the potential benefits to the society from economies of scale of big financial institutions are unlikely to ever exceed the potential costs due to increased risk of financial crisis. Our results have important implications for this debate.

First, our findings suggest that most of the biggest US commercial banks enjoy substantial economies of scale (Obelix is not obese). However, as evidenced in Table 6, not all the banks with assets in excess of $1 billion exhibit economies of scale. Further, the RTS estimates for 35 % of the observations belonging to the top one hundred banks are consistent with constant or decreasing RTS.[32] For the top ten banks with assets ranging from $47 billion to $1.5 trillion, only 70 % of the observations produce RTS estimates consistent with increasing RTS.[33] In particular, of the four banks with assets above $500 billion, on average, only one bank exhibits increasing RTS during the sampling period (only 3 obese Obelix). Thus, capping the size of banks around $1 trillion (converting them to Asterix), for instance, may yield limited social losses from the scale economies viewpoint.

Second, our results indicate that scale economies are likely to continue to be the major driver of growth for small, medium, and some of the biggest banks. We find that scale economies contribute significantly to TFP growth, giving strong incentives for banks to keep growing and, to regulators, a more challenging and difficult task to keep them on lean diet. If regulators want to keep bank size under control, they will have to consider ways in which banks internalize the associated potential cost for society. This will likely require imposing endogenous regulatory constraints that

[32] The banks belonging to the top one hundred banks have assets ranging from $23 million to $1.5 million.

[33] Including the years 2007 and 2008, only 57 % of the observations for the top ten biggest banks are consistent with increasing RTS. Among the top ten biggest banks for each year are State Street Bank and Trust Company, CitiBank, US Bank, Wachovia, HSBC, Wells Fargo, Bank of America, Bank of New York Mellon, Fleet National Bank, Suntrust Bank, Keybank, PNC Bank, Regions Bank, JP Morgan Chase, and Citizens Bank.

increase the marginal cost of getting bigger above the marginal benefits of getting even bigger.

Third, consolidation of small and medium banks and further growth of some of the biggest banks pose big challenges to regulators. As overall bank size increases, widespread bank failures can be more problematic for regulators. Future bank failures will likely involve, on average, bigger and more interconnected banks. Therefore, regulators should widen the focus of their efforts beyond the biggest commercial banks and BHCs and look in the direction of smaller banks that have the strongest incentive to keep growing. In addition, larger average size of banks means stronger barriers to entry for potential competitors which can greatly affect concentration, competition, and efficiency.

7 Conclusions

Regulators in the US and around the world are still pondering the idea of how to keep bank size in check. Thus, understanding whether the biggest US banks enjoy economies of scale is important. In this paper, we present new nonparametric estimates of scale economies, TFP growth and its components (TC, scale, and other environmental variables) for US commercial banks during 2001 to 2010. We find that 73 % of the top one hundred banks, 98 % of medium and small banks, and seven of the top ten biggest banks by asset size exhibit substantial economies of scale.

The existence of substantial scale economies raises an important challenge for bank regulators. Banks achieve lower average costs by scaling up their operations. This, in turn, translates into higher bank profitability. So, any size limit regulation will be naturally challenged by banks. In addition, limiting the size of banks might reallocate banks' activities to other less regulated financial institutions, creating a different regulatory problem.

Our economies of scale estimates are derived from a novel and more flexible approach than those used in previous studies. We start from a nonparametric cost function with fixed bank-specific effects and derive a semiparametric equation with smooth coefficients. The smooth coefficients are nonparametric (fully flexible) functions of the covariates of the cost function and therefore the estimates of scale economies, TFP growth, and its components are also nonparametric. We compare the results from the semiparametric model with two other models that are parametric but flexible. These features are absent from most recent studies.

Our findings suggest that most US commercial banks with assets in excess of $1 billion enjoy substantial economies of scale (Obelix is not obese). However, 35 % of the observations for the top one hundred banks show no evidence of scale economies. For the top ten banks with assets ranging from $47 billion to $1.5 trillion, only 70 % of the observations are consistent with economies of scale. In particular, of the four banks with assets above $500 billion, on average, only one bank exhibits economies of scale during the sampling period (only 3 obese Obelix). Thus, capping the size of the biggest banks (converting them to Asterix) may yield limited social losses from the scale economies viewpoint.

Unlike previous studies, we find that scale economies contribute positively and significantly to bank TFP growth. This support our main conclusions that the US commercial banks still have incentives to growth further – doing so contributes positively and significantly to TFP growth. Our results are robust to alternative model specifications and sampling mechanisms. Our estimates of economies of scale are in line with those presented in the recent banking literature.

Bank TFP grew about 33 % from 2001 to 2010. The main drivers of TFP growth were scale economies and allocative efficiency which account for 21.5 % and 17.3 % of TFP growth, respectively, during this period. The trend of TFP growth for big, medium, and small banks is similar. However, big banks experience sharper swings in TFP growth than medium and small banks. TFP grew steadily for all banks from 2001 to 2010 with some periods of deceleration. During 2008 to 2010, TFP growth for big banks was 1.5 times higher than the TFP growth for medium and small banks.

Overall, our results indicate that the majority of the top one hundred US commercial banks seems to be operating below their optimal size. Thus, further growth and industry consolidation are likely to continue. Specifically, bigger banks may pursue further growth by acquiring small and medium banks which benefit the most from exploiting economies of scale.

Acknowledgments We would like to thank the Editor and two anonymous referees for their helpful and insightful comments. However, the authors are solely responsible for the views expressed in this article. Restrepo-Tobón acknowledges financial support from the Colombian Fulbright Commission; the Colombian Administrative Department of Science, Technology and Innovation (Colciencias); and EAFIT University.

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

8 Appendix 1: TFP growth decomposition

Starting with the standard definition of TFP change in Eq. (7) and adding $T\dot{F}P$ to both sides of Eq. (5) we have:

$$\begin{aligned} &T\dot{F}P + \beta_0(\cdot) + \sum_{k=1}^{K-1} \beta_k(\cdot)\dot{\tilde{W}}_k + \sum_{q=1}^{Q} \gamma_q(\cdot)\dot{Y}_q + \sum_{p=1}^{P} \varphi_p(\cdot)\nabla_t Z_p + u \\ &\equiv \dot{\tilde{C}} + \sum_{q=1}^{Q} R_q \dot{Y}_q - \sum_{k=1}^{K} S_k \dot{X}_k \end{aligned} \tag{10}$$

Using the definitions $\dot{\tilde{C}} = \dot{C} - \dot{W}_K$ and $\dot{C} = \sum_{k=1}^{K} S_k \dot{W}_k + \sum_{k=1}^{K} S_k \dot{X}_k$,the right-hand-side of Eq. (10) can be expressed as:

$$\dot{\tilde{C}} + \sum_{q=1}^{Q} R_q \dot{Y}_q - \sum_{k=1}^{K} S_k \dot{X}_k \equiv \sum_{k=1}^{K} S_k \dot{W}_k + \sum_{q=1}^{Q} R_q \dot{Y}_q - \dot{W}_K \tag{11}$$

Since $\sum_{k=1}^{K} S_k = 1$, $\dot{W}_k = \ddot{\tilde{W}}_k + \dot{W}_K, \forall k = 1, \ldots, K-1$, and $\ddot{\tilde{W}}_K = 0$,

$$\sum_{k=1}^{K} S_k \dot{W}_k + \sum_{q=1}^{Q} R_q \dot{Y}_q - \dot{W}_K \equiv \sum_{k=1}^{K-1} S_k \ddot{\tilde{W}}_k + \sum_{q=1}^{Q} R_q \dot{Y}_q \tag{12}$$

Using this result, the relationship in Eq. (10) can be expressed as:

$$T\dot{F}P \equiv -\beta_0(\cdot) + \sum_{q=1}^{Q} (R_q - \gamma_q(\cdot))\dot{Y}_q + \sum_{k=1}^{K-1} (S_k - \beta_k(\cdot))\ddot{\tilde{W}}_k - \sum_{p=1}^{P} \varphi_p(\cdot)\nabla_t Z_p - u \tag{13}$$

9 Appendix 2: Model specifications and estimation

Following Li et al. (2002) and Li and Racine (2007), the local-constant estimator for $\Psi(z)$ in Eq. (9) is expressed as:

$$\hat{\Psi}(z) = \left[\sum_{i=1}^{N} \sum_{t=1}^{T} \mathcal{X}_{it} \mathcal{X}'_{it} \mathcal{K}\left(\frac{\mathcal{Z}_{it} - z}{h} \right) \right]^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} \mathcal{X}_{it} \mathcal{Y}_{it} \mathcal{K}\left(\frac{\mathcal{Z}_{it} - z}{h} \right) \tag{14}$$

where N and T denotes number of banks and time periods, respectively, h is a $(K + Q + P)$ vector with each element a selected bandwidth for each z variable and $\mathcal{K}(\cdot)$ is the product Gaussian kernel function.[34]

Note that if the kernel function is absent, then the SPSC estimator reduces to its OLS counterpart. The SPSC model also nests the partially linear model proposed by Robinson (1988) as a special case, which makes only the intercept an unknown smooth function of the $\mathcal{Z}$ variables.

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, i.e.,

$$CV_{lc}(h) = \min_{h} \sum_{i=1}^{N} \sum_{t=1}^{T} \left[\mathcal{Y}_{it} - \mathcal{X}'_{it} \hat{\Psi}_{-it}(\mathcal{Z}_{it}) \right]^2 M(\mathcal{Z}_{it}) \tag{15}$$

where $CV_{lc}(h)$ determines the cross-validation bandwidth vector h for local constant estimator, $\mathcal{X}'_{it} \hat{\Psi}_{-it}(\mathcal{Z}_{it})$ is the leave-one-out local-constant kernel conditional mean, and $0 \leq M(\cdot) \leq 1$ is a weight function that serves to avoid difficulties caused by dividing by zero. Unlike other methods proposed in the recent literature (e.g. Feng

[34] Explicitly, the kernel function is written as:

$$\mathcal{K}(\cdot) = \prod_{l=1}^{K+Q+P} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} \left(\frac{\mathcal{Z}_{lit} - z_l}{h_l} \right)^2 \right)$$

and Serletis 2010 and Wheelock and Wilson 2011) the SPG model can be easily estimated using widely available statistical software. For instance, the bandwidths for the $\mathcal{Z}$ variables and the smooth coefficients can be estimated using the *NP* package in R (Hayfield and Racine 2008).[35]

Now, assuming that $f(\cdot)$ in Eq. (3) is translog, the estimating equation for the PL model is:

$$\begin{aligned} f(\cdot) = \alpha_0 &+ \sum_{k=1}^{K-1} \theta_k \ln \widetilde{W}_k + \frac{1}{2} \sum_{k=1}^{K-1} \sum_{m=1}^{K-1} \theta_{km} \ln \widetilde{W}_k \ln \widetilde{W}_m \\ &+ \sum_{q=1}^{Q} \alpha_q \ln Y_q + \frac{1}{2} \sum_{q=1}^{Q} \sum_{o=1}^{Q} \alpha_{qo} \ln Y_q \ln Y_o \\ &+ \sum_{q=1}^{Q} \sum_{k=1}^{K-1} \delta_{qk} \ln Y_q \ln \widetilde{W}_k + \alpha_t\, t + \frac{1}{2} \alpha_{tt}\, t^2 + \sum_{q=1}^{Q} \lambda_{qt} \ln Y_q\, t \\ &+ \sum_{k=1}^{K-1} \delta_{kt} \ln \widetilde{W}_k\, t + \sum_{p=1}^{P} \phi_p Z_p \\ &+ \frac{1}{2} \sum_{p=1}^{P} \sum_{l=1}^{P} \phi_{pl} Z_p Z_l + \sum_{p=1}^{P} \sum_{k=1}^{K-1} \rho_{pk} Z_p \ln \widetilde{W}_k + \sum_{p=1}^{P} \sum_{q=1}^{Q} \tau_{pq} Z_p \ln Y_q \\ &+ \sum_{p=1}^{P} \phi_{pt} Z_p\, t + u \end{aligned} \tag{16}$$

After estimating Eq. (16), the functional coefficients for the PL model are computed using:

$$\begin{aligned} \frac{\partial f}{\partial \ln \widetilde{W}_k} = \beta_k(\cdot) &= \theta_k + \sum_{m=1}^{K-1} \theta_{km} \ln \widetilde{W}_m + \sum_{q=1}^{Q} \delta_{qk} \ln Y_q + \delta_{kt} t \\ &+ \sum_{p=1}^{P} \rho_{pk} Z_p, \ \forall k = 1, \ldots, K-1 \end{aligned} \tag{17}$$

$$\begin{aligned} \frac{\partial f}{\partial \ln Y_q} = \gamma_q(\cdot) &= \alpha_q + \sum_{o=1}^{Q} \alpha_{qo} \ln Y_o + \sum_{k=1}^{K-1} \delta_{qk} \ln \widetilde{W}_k + \lambda_{qt}\, t \\ &+ \sum_{p=1}^{P} \tau_{pq} Z_p, \ \forall q = 1, \ldots, Q \end{aligned} \tag{18}$$

[35] We use the functions *npscoefbw* and *npscoef* with their default values. To decrease the estimation computational time, we set the optimization options to "nmulti = 1", "optim.abstol = 0.000001", and "optim.reltol = sqrt(0.000001)".

$$\frac{\partial f}{\partial Z_p} = \varphi_p(\cdot) = \phi_p + \sum_{l=1}^{P} \phi_{pl} Z_l + \sum_{k=1}^{K-1} \rho_{pk} \ln \widetilde{W}_k + \sum_{q=1}^{Q} \tau_{pq} \ln Y_q + \phi_{pt}\, t,\ \forall p = 1, \ldots, P \tag{19}$$

$$\frac{\partial f}{\partial t} = \beta_0(\cdot) = \alpha_t + \alpha_{tt} t + \sum_{q=1}^{Q} \lambda_{qt} \ln Y_q + \sum_{k=1}^{K-1} \delta_{kt} \ln \widetilde{W}_k + \sum_{p=1}^{P} \phi_{pt} Z_p \tag{20}$$

Plugging Eq. (17)–(20) into Eq. (5) gives:

$$\begin{aligned}
\dot{\tilde{C}} = & \left[\alpha_t + \alpha_{tt} t + \sum_{q=1}^{Q} \lambda_{qt} \ln Y_q + \sum_{k=1}^{K-1} \delta_{kt} \ln \widetilde{W}_k + \sum_{p=1}^{P} \phi_{pt} Z_p \right] \\
& + \sum_{k=1}^{K-1} \left[\theta_k + \sum_{m=1}^{K-1} \theta_{km} \ln \widetilde{W}_m + \sum_{q=1}^{Q} \delta_{qk} \ln Y_q + \delta_{kt} t + \sum_{p=1}^{P} \rho_{pk} Z_p \right] \dot{\widetilde{W}}_k \\
& + \sum_{q=1}^{Q} \left[\alpha_q + \sum_{o=1}^{Q} \alpha_{qo} \ln Y_o + \sum_{k=1}^{K-1} \delta_{qk} \ln \widetilde{W}_k + \lambda_{qt}\, t + \sum_{p=1}^{P} \tau_{pq} Z_p \right] \dot{Y}_q \\
& + \sum_{p=1}^{P} \left[\phi_p + \sum_{l=1}^{P} \phi_{pl} Z_l + \sum_{k=1}^{K-1} \rho_{pk} \ln \widetilde{W}_k + \sum_{q=1}^{Q} \tau_{pq} \ln Y_q + \phi_{pt}\, t \right] \nabla_t Z_p + u
\end{aligned} \tag{21}$$

Rearranging Eq. (21) gives the estimating equation for the PG model as follows:

$$\begin{aligned}
\dot{\tilde{C}} = & \alpha_t + \alpha_{tt} t + \sum_{k=1}^{K-1} \theta_k \dot{\widetilde{W}}_k + \sum_{k=1}^{K-1} \sum_{m=1}^{K-1} \theta_{km} \ln \widetilde{W}_m \dot{\widetilde{W}}_k \\
& + \sum_{q=1}^{Q} \alpha_q \dot{Y}_q + \sum_{q=1}^{Q} \sum_{o=1}^{Q} \alpha_{qo} \ln Y_o \dot{Y}_q \\
& + \sum_{p=1}^{P} \phi_p \nabla_t Z_p + \sum_{p=1}^{P} \sum_{l=1}^{P} \phi_{pl} Z_l \nabla_t Z_p + \sum_{q=1}^{Q} \lambda_{qt} \left(\ln Y_q + t \dot{Y}_q \right) \\
& + \sum_{k=1}^{K-1} \delta_{kt} \left(\ln \widetilde{W}_k + t \dot{\widetilde{W}}_k \right) \\
& + \sum_{p=1}^{P} \phi_{pt} \left(Z_p + t \nabla_t Z_p \right) + \sum_{k=1}^{K-1} \sum_{q=1}^{Q} \delta_{qk} \left(\ln Y_q \dot{\widetilde{W}}_k + \ln \widetilde{W}_k \dot{Y}_q \right)
\end{aligned}$$

$$+\sum_{k=1}^{K-1}\sum_{p=1}^{P}\rho_{pk}\left(Z_p\dot{\widetilde{W}}_k+\ln\widetilde{W}_k\nabla_t Z_p\right)$$
$$+\sum_{q=1}^{Q}\sum_{p=1}^{P}\tau_{pq}\left(Z_p\dot{Y}_q+\ln Y_q\nabla_t Z_p\right)+u \tag{22}$$

After estimating Eq. (22), we compute the functional coefficients for the PG model using Eqs. (17)–(20).

References

Allen, J., & Liu, Y. (2007). Efficiency and economies of scale of large Canadian banks. *Canadian Journal of Economics*, *40*(1), 225–244.

Amel, D., Barnes, C., Panetta, F., & Salleo, C. (2004). Consolidation and efficiency in the financial sector: A review of the international evidence. *Journal of Banking & Finance*, *28*(10), 2493–2519.

Arellano, M., & Honore, B. (2001). Panel data models: Some recent developments. In J. J. Heckman & E. Leamer (Eds.), *Handbook of Econometrics* (chap. 53) (Vol. 5, pp. 3229–3290). Amsterdam: Elsevier.

Bai, J. (2009). Panel data models with interactive fixed effects. *Econometrica*, *77*(4), 1229–1279.

Bauer, P., Berger, A., & Humphrey, D. (1993). Efficiency and productivity growth in US banking. In H. Fried & C. Lovell (Eds.), *The measurement of productive efficiency: Techniques and applications* (pp. 386–413). New York: Oxford University Press. S S.

Berger, A. N., & Humphrey, D. (1994). Bank scale economies, mergers, concentration, and efficiency: The US experience. Center for Financial Institutions Working Papers.

Berger, A. N., & Humphrey, D. B. (1992). Megamergers in banking and the use of cost efficiency as an antitrust defense. *Antitrust Bulletin*, *37*, 541.

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

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

Berger, A. N., Hanweck, G. A., & Humphrey, D. B. (1987). Competitive viability in banking: Scale, scope, and product mix economies. *Journal of Monetary Economics*, *20*(3), 501–520.

Berger, A. N., Demsetz, R. S., & Strahan, P. E. (1999). The consolidation of the financial services industry: Causes, consequences, and implications for the future. *Journal of Banking & Finance*, *23*(2–4), 135–194.

Borts, G. H. (1954). Increasing returns in the railway industry. *The Journal of Political Economy, 62*(3), 316–333.

Boyd, J., & Heitz, A. (2012). The social costs and benefits of too-big-to-fail banks: A "bounding" exercise. Technical report, University of Minnesota.

Camacho, F. T., & Menezes, F. M. (2009). Access pricing and investment: A real options approach. *Journal of Regulatory Economics*, *36*(2), 107–126.

Chamberlain, G. (1984). Panel data. In Z. Griliches & M. Intriligator (Eds.), *Handbook of econometrics* (Vol. 2, pp. 1248–1318). Amsterdam: Elsevier. chap 22.

Clark, J. A. (1988). Economies of scale and scope at depository financial institutions: A review of the literature. *Economic Review*, *73*(8), 17–33.

Clark, J. A. (1996). Economic cost, scale efficiency, and competitive viability in banking. *Journal of Money, Credit and Banking*, *28*(3), 342–364.

Daniels, K. N., & Tirtiroglu, D. (1998). Total factor productivity growth in US commercial banking for 1935–1991: A latent variable approach using the kalman filter. *Journal of Financial Services Research*, *13*, 119–135. doi:10.1023/A:1007922103037.

Das, A., & Kumbhakar, S. (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.

Davies, R., & Tracey, B. (2014). Too big to be efficient? The impact of implicit subsidies on estimates of scale economies for banks. *Journal of Money, Credit and Banking*, *46*(s1), 219–253.
Denny, M., Fuss, M., & Waverman, L. (1979). Productivity measurement in regulated industries, Academic Press, chap The measurement and interpretation of total factor productivity in regulated industries, with an application to Canadian telecommunications (pp. 179–212).
DeYoung, R. (1991). The efficiencies defense and commercial bank merger regulation. *Review of Industrial Organization*, *6*(3), 269–282.
DeYoung, R. (2010). Scale economies are a distraction. *The Region*, *10*(3), 7.
Diewert, W. E., & Fox, K. J. (2008). On the estimation of returns to scale, technical progress and monopolistic markups. *Journal of Econometrics*, *145*(1–2), 174–193.
Ellig, J., & Giberson, M. (1993). Scale, scope, and regulation in the Texas gas transmission industry. *Journal of Regulatory Economics*, *5*(1), 79–90.
Evanoff, D. D., Israilevich, P. R., & Merris, R. C. (1990). Relative price efficiency, technical change, and scale economies for large commercial banks. *Journal of Regulatory Economics*, *2*(3), 281–298.
Evans, L., & Guthrie, G. (2006). Incentive regulation of prices when costs are sunk. *Journal of Regulatory Economics*, *29*(3), 239–264.
Farrell, J., & Shapiro, C. (2001). Scale economies and synergies in horizontal merger analysis. *Antitrust Law Journal*, *68*(3), 685–710.
Feldman, R. (2010). Size and regulatory reform in finance: Important but difficult questions. *The Region, Federal Reserve Bank of Minneapolis*, *10*, 7.
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. *Journal of Applied Econometrics*, *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. *Journal of Banking & 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. *Journal of Banking & Finance*, *36*(6), 1883–1895.
Foreman, R. D., & Beauvais, E. (1999). Scale economies in cellular telephony: Size matters. *Journal of Regulatory Economics*, *16*(3), 297–306.
Fraquelli, G., Piacenza, M., & Vannoni, D. (2005). Cost savings from generation and distribution with an application to Italian electric utilities. *Journal of Regulatory Economics*, *28*(3), 289–308.
Gandhi, P., & Lustig, H. (In press). Size anomalies in US bank stock returns. Journal of Finance, pp. 1–39.
Glass, V., & Stefanova, S. K. (2012). Economies of scale for broadband in rural United States. *Journal of Regulatory Economics*, *41*(1), 100–119.
Greene, W. (2005a). Fixed and random effects in stochastic frontier models. *Journal of Productivity Analysis*, *23*, 7–32. doi:10.1007/s11123-004-8545-1.
Greene, W. (2005b). Reconsidering heterogeneity in panel data estimators of the stochastic frontier model. *Journal of Econometrics*, *126*(2), 269–303.
Hardle, W., & Mammen, E. (1993). Comparing nonparametric versus parametric regression fits. *The Annals of Statistics*, *21*(4), 1926–1947.
Hayfield, T., & Racine, J. S. (2008). Nonparametric econometrics: The np package. *Journal of Statistical Software*, *27*(5), 1–32.
Hughes, J. P., & Mester, L. J. (1998). Bank capitalization and cost: Evidence of scale economies in risk management and signaling. *Review of Economics and Statistics*, *80*(2), 314–325.
Hughes, J. P., & Mester, L. J. (2010). Efficiency in banking: Theory, practice, and evidence. In A. Berger, P. Molyneux, & J. Wilson (Eds.), *The oxford handbook of banking* (chap. 19) (pp. 336–357). New York: Oxford University Press.
Hughes, J. P., & Mester, L. J. (2013). Who said large banks don't experience scale economies? Evidence from a risk-return-driven cost function. *Journal of Financial Intermediation*, *22*(4), 559–585.
Hughes, J. P., Lang, W., Mester, L. J., & Moon, C. G. (1996). Efficient banking under interstate branching. *Journal of Money, Credit and Banking*, *28*(4), 1045–1071.
Hughes, J. P., Lang, W., Mester, L. J., & Moon, C. G. (2000). Recovering risky technologies using the almost Ideal demand system: An application to US banking. *Journal of Financial Services Research*, *18*, 5–27.

Hughes, J. P., Mester, L. J., & Moon, C. G. (2001). Are scale economies in banking elusive or illusive?: Evidence obtained by incorporating capital structure and risk-taking into models of bank production. *Journal of Banking & Finance*, *25*(12), 2169–2208.
Humphrey, D. (1992). Flow versus stock indicators of banking output: Effects on productivity and scale economy measurement. *Journal of Financial Services Research*, *6*(2), 115–135.
Humphrey, D. B. (1991). Productivity in banking and effects from deregulation. *Economic Review*, *1*(Mar), 16–28.
Humphrey, D. B. (1993). Cost and technical change: Effects from bank deregulation. *Journal of Productivity Analysis*, *4*(1/2), 9–34.
Hunter, W. C., & Timme, S. G. (1986). Technical change, organizational form, and the structure of bank production. *Journal of Money, Credit and Banking*, *18*(2), 152–166.
Hunter, W. C., & Timme, S. G. (1991). Technological change in large US commercial banks. *The Journal of Business*, *64*(3), 339–362.
Johnson, S. (2012). Tarullo telegraphs Fed's plans to cap bank size. Bloomberg News' Column. http://goo.gl/Wr4Y3M
Kinne, K. (1998). The "efficiency defense" in the US American merger policy. Technical report, HWWA Discussion Paper.
Kolasky, W. J., & Dick, A. R. (2003). The merger guidelines and the integration of efficiencies into antitrust review of horizontal mergers. *Antitrust Law Journal*, *71*, 207–251.
Kumbhakar, S., & Lozano-Vivas, A. (2005). Deregulation and productivity: The case of Spanish banks. *Journal of Regulatory Economics*, *27*(3), 331–351.
Kumbhakar, S., & Sun, K. (2012). Estimation of TFP growth: A semiparametric smooth coefficient approach. *Empirical Economics*, *43*(1), 1–24.
Kumbhakar, S., Lien, G., Flaten, O., & Tveters, R. (2008). Impacts of Norwegian milk quotas on output growth: A modified distance function approach. *Journal of Agricultural Economics*, *59*(2), 350–369.
Kumbhakar, S. C., & Wang, D. (2007). Economic reforms, efficiency and productivity in Chinese banking. *Journal of Regulatory Economics*, *32*(2), 105–129.
Lagerlof, J. N., & Heidhues, P. (2005). On the desirability of an efficiency defense in merger control. *International Journal of Industrial Organization*, *23*(9–10), 803–827.
Laudati, L. L. (1981). Note: Economies of scale: Weighing operating efficiency when enforcing antitrust law. *Fordham Law Review*, *49*, 771–801.
Li, Q., & Racine, J. (2007). *Nonparametric econometrics: Theory and practice*. Princeton: Princeton University Press.
Li, Q., & Racine, J. S. (2010). Smooth varying-coefficient estimation and inference for qualitative and quantitative data. *Econometric Theory*, *26*(06), 1607–1637.
Li, Q., Huang, C. J., Li, D., & Fu, T. T. (2002). Semiparametric smooth coefficient models. *Journal of Business & Economic Statistics*, *20*(3), 412–422.
Malikov, E., Restrepo-Tobón, D., Kumbhakar, S.C. (2014). Estimation of banking technology under credit uncertainty. *Empirical Economics* Forthcoming, 1–27.
McAfee, R. P., Mialon, H. M., & Williams, M. A. (2004). What is a barrier to entry? *American Economic Review*, *94*(2), 461–465.
Mester, L. J. (1994). How efficient are third district banks? *Business Review*, 1–3.
Mester, L. J. (1996). A study of bank efficiency taking into account risk-preferences. *Journal of Banking & Finance*, *20*(6), 1025–1045.
Mester, L. J. (1997). Measuring efficiency at US banks: Accounting for heterogeneity is important. *European Journal of Operational Research*, *98*(2), 230–242.
Mukherjee, K., Ray, S. C., & Miller, S. M. (2001). Productivity growth in large US commercial banks: The initial post-deregulation experience. *Journal of Banking & Finance*, *25*(5), 913–939.
Nauges, C., & Van Den Berg, C. (2008). Economies of density, scale and scope in the water supply and sewerage sector: A study of four developing and transition economies. *Journal of Regulatory Economics*, *34*(2), 144–163.
Restrepo-Tobón, D., Kumbhakar, S. (2013). Profit efficiency of US commercial banks: A decomposition. Technical report 13–18, EAFIT University, Binghamton University.
Restrepo-Tobón, D., & Kumbhakar, S. (2014). Nonparametric estimation of returns to scale using input distance functions: An application to large US banks. *Empirical Economics*, *48*(1), 143–168.
Robinson, P. (1988). Root-n-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society*, *56*(4), 931–954.

Rogers, K. (1998). Nontraditional activities and the efficiency of US commercial banks. *Journal of Banking & Finance*, *22*(4), 467–482.

Schmalensee, R. (2004). Sunk costs and antitrust barriers to entry. *American Economic Review*, *94*(2), 471–475.

Sealey, C., & Lindley, J. (1977). Inputs, outputs, and a theory of production and cost at depository financial institutions. *Journal of Finance*, *32*(4), 1251–1266.

Semenick Alam, I. M. (2001). A nonparametric approach for assessing productivity dynamics of large US banks. *Journal of Money, Credit and Banking*, *33*(1), 121–139.

Shaffer, S. (1994). A revenue-restricted cost study of 100 large banks. *Applied Financial Economics*, *4*(3), 193–205.

Stern, G., & Feldman, R. (2009 June). Addressing TBTF by shrinking financial institutions: An initial assessment. Federal Reserve Bank of Minneapolis: The Region (pp. 8–13).

Stiroh, K. J. (2000). How did bank holding companies prosper in the 1990s? *Journal of Banking & Finance*, *24*(11), 1703–1745.

Tirtiroglu, D., Daniels, K. N., & Tirtirogu, E. (2005). Deregulation, intensity of competition, industry evolution, and the productivity growth of US commercial banks. *Journal of Money, Credit and Banking*, *37*(2), 339–360.

Wang, H. J., & Ho, C. W. (2010). Estimating fixed-effect panel stochastic frontier models by model transformation. *Journal of Econometrics*, *157*(2), 286–296.

Wheelock, D. C., & Wilson, P. W. (1999). Technical progress, inefficiency, and productivity change in US banking, 1984–1993. *Journal of Money, Credit and Banking*, *31*(2), 212–234.

Wheelock, D. C., & Wilson, P. W. (2001). New evidence on returns to scale and product mix among US commercial banks. *Journal of Monetary Economics*, *47*(3), 653–674.

Wheelock, D. C., & Wilson, P. W. (2011). Are credit unions too small? *Review of Economics and Statistics*, *93*(4), 1343–1359.

Wheelock, D. C., & Wilson, P. W. (2012). Do large banks have lower costs? New estimates of returns to scale for US banks. *Journal of Money, Credit and Banking*, *44*(1), 171–199.

Williamson, O. E. (1968). Economies as an antitrust defense: The welfare tradeoffs. *American Economic Review*, *58*(1), 18–36.

Williamson, O. E. (1977). *Welfare aspects of industrial markets* (pp. 237–271) (Economies as an antitrust defense revisited). Leiden: Springer.