Journal of Productivity Analysis
https://doi.org/10.1007/s11123-019-00565-6

A dynamic stochastic frontier model with threshold effects: U.S. bank size and efficiency

Pavlos Almanidis[1] · Mustafa U. Karakaplan[2] · Levent Kutlu[3]

Abstract

Common/Single frontier methodologies that are used to analyze bank efficiency and performance can be misleading because of the homogeneous technology assumption. Using the U.S. banking data over 1984-2010, our dynamic methodology identifies a few data-driven thresholds and distinct size groups. Under common frontier assumption, the largest banks appear to be 22% less efficient on average than how they are in our model. Also, in the common frontier model, smaller banks seem to be relatively more efficient compared to their larger counterparts. Hence, common policies or regulations may not be well-balanced about controlling the banks of different sizes on the spectrum.

Keywords Dynamic Stochastic Frontier · Bank Efficiency · Bank Heterogeneity

JEL classification C13 · C23 · D24 · G21 · G28

1 Introduction

The overall landscape of the U.S. banking sector changed tremendously over the last three decades. Major deregulatory acts, such as the Riegle-Neal Interstate Banking and Branching Efficiency Act of 1994 allowed interstate banking and branching, and the Gramm–Leach–Bliley Financial Services Modernization Act of 1999 granted broad-based securities, investment, and insurance power to commercial banks. Industrial and digital/electronic improvements, as well as the introduction of new financial products revolutionize the banking sector. Automated teller machines, internet banking, electronic payments and processing, and the emergence of new complex financial products such as collateralized mortgage obligations and credit default swaps were among the few developments that allowed banks to transform from traditional intermediary banks of "originate-and-hold" into complex multi-level entities. Banking crises, on the other hand, took their toll in this period on most banking units operating in the U.S. banking industry. Combined with the competitive pressures from the new market entrants, the banking crises such as the savings and loan crisis of the 1980s and 1990s and the latest 2007–2010 banking crunch eliminated more than 8,000 banks over the period under consideration. These changes and developments in the banking industry raise questions about the efficiency and performance of the U.S. banks and the nature and effectiveness of regulations that are intended to control them.

Traditional common frontier models are frequently used to evaluate bank efficiency and performance. What is often overlooked with these models, however, is that the banking sector may be characterized by heterogeneous technologies/frontiers[1] and assuming a single technology for the whole

Supplementary information The online version of this article (https://doi.org/10.1007/s11123-019-00565-6) contains supplementary material, which is available to authorized users.

✉ Levent Kutlu
levent.kutlu@utrgv.edu

1 International Tax Services, Transfer Pricing, Ernst & Young LLP, Ernst & Young Tower, 100 Adelaide Street West, PO Box 1, Toronto, ON M5H 0B3, Canada

2 College of Business, Governors State University, University Park, IL 60484, USA

3 Department of Economics and Finance, University of Texas Rio Grande Valley, 1201 W. University Dr., Edinburg, TX 78539, USA

[1] Throughout this paper, whenever we refer to "technology", we mean the production possibilities frontier (or simply the frontier) corresponding to that particular technology. Hence, at a given time period, if two groups of banks have different frontiers, we say that these groups of banks have different production technologies. See Färe and Primont (1990) for relevant discussions.

banking sector may potentially lead to inconsistent parameter estimates and distorted efficiency rankings. There are several reasons why banks may have different technologies and the implication is that production possibilities can be different for each bank over time. In this paper, in order to have a better understanding of the efficiency and performance of the U.S. banks, we use a dynamic stochastic frontier methodology that allows for bank heterogeneity. In particular, we are interested in examining the technical efficiency of the banks, which measures how successful the banks are in the production process given their resources and the available technology. One important feature of our dynamic model is that group specific technological progress can be captured with a flexible data-driven approach.

Using a quarterly unbalanced panel data from the U.S. banking sector over the 1984–2010 period, we identify that six size-groups exist in the sector, and we estimate time-varying bank-specific and group-specific efficiency scores and scale elasticity estimates. We find that on average, the two top-tier size groups of banks were more efficient than the smaller banks over the sample period. We also show that estimating the model under the assumption of a common frontier results in considerably lower average efficiency scores for larger banks. Moreover, we provide evidence that smaller banks appear to be more efficient under the common frontier assumption than they are in our dynamic model. These results indicate that actual performances of the U.S. banks can be immensely different than what static measures based on a single frontier would predict. Hence, policy implications and regulations based on such measures may be stricter or looser than necessary, or controlling one end of the size spectrum while ignoring the other end.

2 Background and motivation

2.1 Analyses of bank efficiency and performance

Government departments such as the U.S. Department of Labor and the U.S. Department of Commerce use methods based on labor productivity to measure the productivity of banks. In addition to these methods, there are many studies in the literature analyzing bank efficiency and performance using frontier efficiency methodologies such as stochastic frontier analysis, thick frontier analysis, and data envelopment analysis. A common theme in these studies is an investigation of the relationship between bank efficiency/performance and deregulation, mergers and acquisitions, or bank ownership.

For example, using a data envelopment approach and the Malmquist Index, Wheelock and Wilson (1999) present that the technical efficiency and average productivity of the U.S. banks declined over 1984-1993 due to the changes in technology, regulations, and market competition. Mukherjee et al. (2001) conduct a data envelopment analysis to examine the U.S. banks between 1984 and 1990. They find that deregulation improved the efficiency and productivity of the U.S. banks significantly. Similarly, Alam (2001) uses a data envelopment methodology and the Malmquist Index to study the U.S. bank productivity from 1980 to 1989 and shows that the productivity improvements in the 80s were mostly due to the shift in technology instead of changes in efficiency. Berger and Mester (2003) apply a frontier methodology to evaluate the cost and profit productivities of the banks in the U.S in the 1984–1997 period. They present that banks, and especially those involved in mergers, improved their profit productivity and worsened their cost productivity between 1991 and 1997. With a data envelopment methodology, Mehdian et al. (2007) study the U.S. commercial banks over the period 1990–2003 and show that globalization and financial deregulation influence a deterioration in the overall efficiency of the banks.

In addition to the studies based on U.S. banks, many researchers such as Kraft and Tırtıroğlu (1998), Lang and Welzel (1999), Battese et al. (2000), Christopoulos and Tsionas (2001), Isik (2007), Kumbhakar and Wang (2007), and others utilize various frontier methodologies to investigate the efficiency and productivity of banks in other countries. More recent studies such as Berger et al. (2009, 2010), and Lin and Zhang (2009) provide evidence that state-owned banks are less efficient compared to joint-stock banks in China. Lastly, Chang et al. (2012) introduces a productivity index to measure the sources of bank productivity growth in China between 2002 and 2009 and present that the technical progress in capital productivity is the major factor influencing the productivity improvement and total factor technical change.[2]

After the most recent financial crisis, the size and efficiency of the banks took center stage once again in the worldwide discussion of financial matters. Many researchers such as Demirgüç-Kunt and Huizinga (2013), Brewer and Jagtiani (2013), Oliveira et al. (2015) and others investigate the nature and potential perils of the too-big-to-fail banks using local and international data. In the U.S., while the debate on financial reform had been going on in the Senate, serious attempts were made to shrink the largest banks by imposing a size constraint. The SAFE Banking Act, or so-called Brown-Kaufman amendment, was proposed in the Senate in 2010. Many Republicans as well as some Democrats and President Obama's administration

[2] For other banking studies that examine efficiencies include Berger and Humphrey (1997), Vivas (1997), Muñiz (2002), Brissimis et al. (2010), Paradi et al. (2012), Galán et al. (2015), Dong et al. (2016), Tsionas (2017), and Delis et al. (2017).

opposed the proposal, and it failed on the Senate floor by a vote of 61 to 33. The "if size matters or not" discussion, however, did not end by this failure, and the debate is actually still hot. Looking at the past and current political and financial environment in the U.S., we feel that it is important to focus on U.S. banking industry for a further investigation of the subject matter.

2.2 Banking sector in a dynamic frontier framework

Originated by Aigner et al. (1977), Meeusen and van den Broeck (1977), and Battese and Corra (1977), stochastic frontier models provide the fundamentals for myriads of theoretical and applied work that attempt to measure the performance of individuals, firms and countries. Stochastic frontier models assume the existence of a parametric functional form, such as Cobb-Douglas, translog etc., which governs the relationship between the dependent and independent variables. The error term in stochastic frontier models is composed of two parts: a two-sided error term, which captures random shocks, statistical noise and other measurement errors; and a one-sided error term, which captures the effects of inefficiencies relative to the stochastic frontier. Both error terms are assumed to be independent and identically distributed random variables, independent of each other and of the right-hand-side variables.

The traditional frontier models assume a common technology/frontier for all firms in the sample and estimate individual-specific efficiency scores relative to the performance of the best-practice firm, which is assumed as perfectly efficient. Assumption of a common technology is not unreasonable for relatively homogeneous samples of firms, but is unduly restrictive if firms employ different technologies within the same industrial sector. In the banking sector, and other industries that are possibly characterized by heterogeneous technologies, estimating a stochastic frontier model under a common frontier assumption may potentially lead to inconsistent parameter estimates and distorted efficiency rankings.[3]

In practice, access to the best-practice technology is not instantly available to all banks and there are several reasons why banks may have different technologies. For example, there may exist certain restrictions such as size, cash flow and liquidity, geography of operations, and other managerial and regulatory factors that prevent banks from adopting the best-practice technology. In cases where such restrictions do not exist or do not prevent banks from accessing new technologies, there may still be considerable delays in adopting new technologies due to costs and the overall implementation process. According to the argument in Huang (2004), when costs related to the installation and personal training differ across banks, there would be a high variation in technology at any given time. That is, a bank can find it optimal to employ its existing technology if the costs of obtaining and switching to a new technology significantly exceed the benefits derived from the new and potentially superior production technology. For example, while it may be more profitable for some banks to use computers and the internet, other banks may implement a more in-person approach and this approach may be simply due to the new technology's being too costly for them to implement. This situation may also partially explain the lags in acquiring and adopting the new technologies observed among banks operating within the same sector.

Berger and Mester (2003) explore some possibilities of why there would be heterogeneities in the banking industry. They explain that small and large banks use different technologies to create different products which cannot be distinguished in the balance sheet entries. As in their example, a \$1 billion loan issued by a large bank can be considered as a different product requiring different approaches than 10,000 loans of \$100,000 issued in total by small banks. They also mention that large and small banks adopt technologies at different rates and adapt to regulations and deregulations in different ways, both of which would result in heterogeneous banks. Other potential explanations that they investigate include revenue-based productivity gains, level of risk-taking, degree of conventional market power, bank mergers, and bank entry and exit, all of which would result in heterogeneities within the banking industry in general.

2.3 Motivation

If the best-practice technology is state dependent, different production frontiers may not necessarily be aggregated into a common frontier[4] The implication is that production possibilities frontier can be different for each bank in the industry over time, and essentially, it is the assumption of common frontier that would require justification instead of relaxing that assumption to allow for producer heterogeneity. Based on the arguments above, we see a need for relaxing the assumption of a common technology/frontier when analyzing our sample of U.S. commercial banks. One approach to account for heterogeneous technologies in the stochastic frontier framework is through the threshold effects model proposed by Hansen (1999, 2000). In that model, observations are assigned into discrete class-regimes based on consistently estimated cut-off points of an

[3] See Tsionas (2002), El-Gamal and Inanoglu (2005), Greene (2005), and Almanidis (2013) for a few examples.

[4] This means that the econometric issue in estimating the best-practice technologies may not necessarily be solved by simply increasing the flexibility of the functional forms or using non-parametric methods that ignore state dependence.

exogenous threshold variable. For instance, if the threshold variable is the total asset size and there is a single cut-off point, then the observations would be divided into two groups based on their respective value of the total assets at a particular point of time. With this methodology, different observations of a bank can be classified into different groups depending on the observational preconditions.

In this paper, we extend the non-dynamic stochastic frontier model considered by Almanidis (2013)[5] to a setting in which the one-sided error term is dynamic, so that our model can capture technological changes in a more flexible way. In this model, banks can switch their technology groups based on the exogenous variables that determine their technology/frontier such as their size, resources, and products. This possibility captures the idea that banks adopt technologies at different rates as discussed in Berger and Mester (2003), and over time, some banks can get ahead or fall behind the technological improvements and innovations. A similar technological shift may happen when two banks merge into a larger bank, and share and jointly improve their technologies. This is a reasonable assumption especially when the time period under the study is long. Our model allows this type of technological changes, which in turn allows group specific production technologies to vary at different rates.

Unlike Berger and Mester (2003) and others[6], we do not arbitrarily classify banks into a number of size groups, but rather use statistical tools to determine what the number of bank groups and the corresponding asset sizes that determines the technology level should be. It is a common practice to categorize banks with respect to arbitrary thresholds. The threshold levels that are used to determine how big is too big and designate Systemically Important Financial Institution (SIFI) status of the Dodd-Frank act are often criticized for not having a strong analytical foundation. Also, using fixed thresholds for regulations can create incentives for banks to operate according to these fixed thresholds to evade regulations. Even the former Representative Barney Frank (2015) suggested changing the original threshold levels of his namesake law to an index that is not predetermined. Hence, we do not classify banks arbitrarily as small or large[7], but instead, offer a statistical procedure to estimate the size-groups by utilizing the bootstrapping algorithm of Hansen (1999, 2000).

Our main data sample includes U.S. commercial banks with average asset sizes of $1 billion or more and excludes the smallest community banks with average asset sizes less than $1 billion. The average asset size of the largest banks in our sample is about 200 times more than that of the smallest banks in our sample. That is, even though community banks are excluded from our main sample, the variation in size (and in other variables) is still substantial, which yields the question of whether it is sensible to assume the same technology for the banks in our main sample. We still expect to observe heterogeneities in our sample and identify different bank groups. One advantage of our methodology is that, within the selected data sample, if the dynamic algorithm cannot discover substantially different technologies based on the structure of the variables integrated in the model, then the number of threshold levels would go down potentially to zero. Such an outcome would support the use of a common frontier methodology. However, if our methodology identifies several threshold levels, that would justify the existence of different technologies and bank groups within the sample. Yet, we also analyze the sample with the restrictive single frontier assumption to compare and contrast the efficiency scores and rankings from different bank groups to those obtained with the aggregated model, and discuss which model captures the financial phenomena better. Furthermore, we examine the data from the community banks that was originally excluded from the main analyses to assess the sensitivity of our main findings.

To sum up, the main objectives of our study are to examine heterogeneities in the U.S. banking sector, identify data-driven threshold levels to categorize banks, and pinpoint different efficiency trends in different bank groups that can be explained by the effects of major deregulatory reforms, banking crises, as well as financial, industrial and digital/electronic improvements and innovations during the sample period that spans from the first quarter of 1984 to the second quarter of 2010. In addition, our goals include comparing the efficiencies of bank groups with each other, identifying the most efficient banks in the industry, and discussing the suitability of the past and current banking regulations and policy proposals based on our findings.

[5] Alternative approaches for dealing with heterogeneous technologies in the stochastic frontier models were considered by Tsionas (2002), Huang (2004), and Greene (2005) in a random parameters model framework, as well as the latent class model employed by Orea and Kumbhakar (2004), El-Gamal and Inanoglu (2005), and Greene (2005). These methodologies require the parameters to be re-estimated with new information and this update may become computationally expensive, especially when the sample size is relatively large. In contrast, the threshold effect model's parsimonious setup would allow the individual units to switch groups, as a result of change in certain characteristics such as size, without requiring a re-estimation of the parameters.

[6] There are many different size groups defined and used by banking researchers and authorities. The cutoff points in almost all cases are arbitrarily selected.

[7] In this paper, whenever we refer to a bank as "small" or "large", this reference would be vaguely saying that the bank belongs to one of the bank groups in our sample that has relatively smaller or larger asset sizes compared to the other groups in our sample. Hence, in this context, even though a bank that has $1 billion asset size may be considered as large compared to smaller community banks in the sector, we consider this bank as a relatively small bank compared to the other banks in our sample with much larger asset sizes.

3 Model and application

Our model is an extension of the model of Duygun et al. (2016) (DKS), which employs the Kalman (1960) filtering techniques[8] to estimate firm efficiencies in the panel data framework. In contrast to their model, our model employs a second degree local approximation to the effects term, and collapses to the random effects model of Cornwell et al. (1990) (CSS) when the variance of the innovation term for the transition equation that determines the effects term is zero. We allow the local approximation to vary over time as in the local level model[9] which enables us to capture time variations of group specific frontiers in a flexible way. In contrast to the model of DKS, our model allows group specific frontiers which are determined by thresholds parameters that are obtained through a bootstrapping procedure. Appendix A provides the details of our local CSS (LCSS) estimator with threshold effects.

Estimation of the threshold parameters is performed by maximizing the log-likelihood function presented in Appendix B using a grid search algorithm on a threshold variable. Values of the threshold variable that maximize the log-likelihood function constitute the solution to the optimization algorithm. One limitation of this method is that the cut-off point parameters of a threshold variable are not identified under the null hypothesis of no threshold(s). As a result, the asymptotic distributions of threshold parameters and the related test statistics have non-standard forms under the null hypothesis. The non-standard test statistic distributions consequently complicate inference. This problem is referred to as Davies' Problem (1987) in the econometrics literature. In order to overcome this limitation, we follow (Hansen 1996, 1999) and employ a bootstrap method presented in Appendix C to simulate the asymptotic distribution of the likelihood ratio test, which is subsequently used in hypothesis testing and to construct the respective confidence intervals.

3.1 Data and variables

The banking data that we examine in this study come from the quarterly consolidated reports of condition and income (Call Reports)[10], which are collected and administrated by the Federal Deposit Insurance Corporation (FDIC) and are collectively available from the Federal Reserve Bank of Chicago's website.[11] We merged the quarterly Call Reports between the first quarter of 1984 and the second quarter of 2010 to create an unbalanced sample of banks. All nominal values were subsequently adjusted to inflation to reflect the values with respect to the Consumer Price Index of 2000.

For our main estimations, we dropped the banks with less than \$1 billion average real asset size over time (community banks), and concentrated on the banks with \$1 billion or more average real asset size over time. This decision was mostly due to the fact that including community banks in the estimation sample would make the application of the LCSS model a relatively complicated time-demanding computational exercise as the bulk of the data comes from those banks.[12] Nevertheless, the banks with \$1 billion or more average real asset size over time represent the major portion of the U.S. commercial banking industry's total asset size and revenues. Even though community banks are out of our main sample, due to the variation within the sample, we still expect to observe heterogeneities in our sample and identify different bank groups. Nevertheless, we also analyzed the data from the community banks under the common frontier assumption to examine the sensitivity of our main findings.

Topics such as proper specification of outputs and inputs and the choice of production or cost functional forms are long-standing controversies of the empirical banking literature.[13] We adopt the intermediation approach of Sealey and Lindley (1977) in our application, which is widely employed in studies of banking performance.[14] According to the intermediation approach, banks are viewed as financial intermediaries that collect deposits and other funds and transform them into loanable funds and other earning assets by using capital and labor. This approach views deposits as inputs, which are treated as outputs in the production and value-added approaches.[15]

Table 1 presents the summary statistics of the variables that we use. Total assets of a bank (K), which is the sum of cash, securities, total loans and leases, and other tangible and intangible assets, is commonly used in the literature as a proxy to measure the size of banks in the banking sector.[16] So we choose this variable as the threshold determinant of

[8] Durbin and Koopman (2012) provide details on the Kalman filter estimation techniques.

[9] Other examples to the models where the one-sided error term is dynamic include Hultberg et al. (2004) and Ahn and Sickles (2000).

[10] The Call Reports contain detailed data on a bank's on-balance and off-balance sheet assets and liabilities, capital structure, income from earning assets, expenses, and other bank-specific structural and geographical characteristics.

[11] www.chicagofed.org

[12] It is worth mentioning that, the number of unique community banks over the sample period exceeds 11,000 and their total number of observations is more than 750,000.

[13] See DeYoung (2013) for more discussion.

[14] See Kaparakis et al. (1994), Wheelock and Wilson (1995, 2001), Sickles (2005), Greene (2005), Kutlu (2012), and Almanidis (2013) among others.

[15] See Baltensperger (1980) and Berger and Humphrey (1992) for the production and value-added approaches discussions.

[16] Note that, although the total assets variable is a commonly employed threshold variable in the banking literature and in practice to assign banks in group sizes, there are other variables/factors that can also be employed separately or in combination to segment the banking industry. Such variables/factors may include a strategy mix, marketing, risk-taking, etc. A proper quantification of these factors in a multi-threshold model framework would potentially complete another piece of the banking technology/strategy puzzle.

Table 1 . Summary Statistics of the Single-Group

	Variable	Mean	Standard deviation	Minimum	Maximum
K	Total assets	\$11.80 B	\$56.60 B	\$8,788.3 T	\$1430.0 B
Y_1	Real estate loans	\$3.00 B	\$14.20 B	\$103.7 T	\$383.0 B
Y_2	Commercial and industrial loans	\$1.92 B	\$7.54 B	\$2.3 T	\$150.0 B
Y_3	Loans to individuals	\$1.02 B	\$4.64 B	\$23.3 T	\$126.0 B
Y_4	Securities	\$1.80 B	\$10.40 B	\$0.9 T	\$346.0 B
Y_5	Off-balance sheet items	\$5.16 B	\$74.90 B	\$28.2 T	\$8430.0 B
X_1	Demand deposits	\$1.10 B	\$3.43 B	\$128.8 T	\$85.8 B
X_2	Total time and saving deposits	\$6.99 B	\$33.00 B	\$1,852.7 T	\$811.0 B
X_3	Labor	3015	11,045	1	215,670
X_4	Capital	\$0.13 B	\$0.46 B	\$16.1 T	\$9.4 B
X_5	Purchased funds	\$2.03 B	\$6.35 B	\$146.1 T	\$148.0 B
N	Number of observations = 36,174				

B and T stand for billions and thousands

the bank size groups. The mean and median of this variable in our sample is \$11.8 billion and \$2.40 billion, respectively, with a minimum of \$8.79 million and a maximum of \$1.43 trillion.

Consistent with Kaparakis et al. (1994), Adams et al. (1999) and Greene (2005), we consider banks as producing and providing five outputs: Real estate loans that are made to individual and businesses secured by real estate (Y_1). Commercial and industrial loans that are made with commercial or industrial purposes to business enterprises (Y_2). Loans to individuals including loans for automobiles, household appliances or furniture, education, medical expenses, and other miscellaneous personal loans and credit cards (Y_3). Securities including U.S. Treasury securities, U. S. government agency and corporation obligations, securities issued by states and political subdivisions in the U.S., and other domestic and foreign debt and equity securities (Y_4). Finally, the off-balance sheet items such as loan commitments, letters of credit, investment structured vehicles, and derivative contract including futures and forwards, swaps, and options (Y_5).[17]

In addition to the five output variables, we use five bank input variables: Demand deposits, which is the total noninterest-bearing transaction accounts (X_1). Total time and savings deposits, which is the sum of interest-bearing savings deposits accounts (X_2). Labor, which is equal to the number of full-time equivalent employees employed by the reporting bank (X_3). Capital including equipment, furniture, fixtures, and capitalized leases (X_4). Lastly, the purchased funds such as wholesale certificate of deposits, federal funds purchased and all securities sold under agreements to repurchase, other borrowed money and notes issued to the U.S. Treasury, brokered deposits, and subordinated notes and debentures (X_5).

After dropping bank observations with zero output and input levels, as well as those with obvious measurement errors and inconsistencies, the final data set includes 573 banks with a total of 36,174 observations.

3.2 Empirical method

Following Coelli and Perelman (1999, 2000) and O'Donnell and Coelli (2005), we consider a multi-input and multi-output translog stochastic output distance frontier.[18] We impose homogeneity of degree one in outputs and add the relevant error terms, namely the symmetric error term $\varepsilon_{i,t}$ and the one-sided inefficiency $u_{i,t}\geq 0$. After rearranging the output distance function, we get:

$$\begin{aligned} y^*_{1,i,t} &= \sum_{m=2}^{M} \alpha_m\,(k_{i,t})\,y^*_{m,i,t} + \sum_{j=1}^{J} \beta_j\,(k_{i,t})\,x_{j,i,t} + \tfrac{1}{2}\sum_{m=2}^{M}\sum_{n=2}^{M} \alpha_{mn}\,(k_{i,t})\,y^*_{m,i,t}y^*_{n,i,t} \\ &+ \tfrac{1}{2}\sum_{j=1}^{J}\sum_{l=1}^{J} \beta_{jl}\,(k_{i,t})\,x_{j,i,t}x_{l,i,t} + \sum_{j=1}^{J}\sum_{m=2}^{M} \delta_{jm}\,(k_{i,t})\,x_{j,i,t}y^*_{m,i,t} + \mu_{i,t} + \varepsilon_{i,t} \end{aligned} \tag{1}$$

where $Y_{m,i,t}$ is the m^{th} output variable and $X_{j,i,t}$ is the j^{th} input variable for the bank i at time t; $\alpha(k_{i,t})$, $\beta(k_{i,t})$, and $\delta(k_{i,t})$ are group specific parameters describing the production technology; $(k_{i,t})$ is the variable that is determining the group that bank i belongs to at time t; $y^*_{1,i,t} = \ln(1/Y_{1,i,t})$; $y^*_{m,i,t} = \ln(Y_{m,i,t}/Y_{1,i,t})$; $x_{j,i,t} = \ln X_{j,i,t}$; $x_{l,i,t} = \ln X_{l,i,t}$; and $\mu_{i,t} = \alpha_{0t}(k_{i,t}) + u_{i,t}$ is the intercept for the bank i at time

[17] In our view, these output and input variables cover the majority of outputs and inputs produced by an average bank operating in the U.S. commercial banking industry.

[18] A translog function provides the second-order Taylor series approximation to any arbitrary function at a single point. In addition, the translog output distance function does not restrict the returns to scale measures and factor demand elasticities to be constant, as is required in the Cobb-Douglas output distance function case.

t. We use real estate loans as the numeraire output variable in our estimation process.

The sum of the absolute values of partial input elasticities $(\sum_j |\partial \ln D_{i,t} / \partial \ln X_{j,i,t}| \geq 0)$ provides a measure of scale elasticities under the output distance function specification. A bank would display increasing, constant, or decreasing returns to scale in a given time period if this measure is greater, equal, or less than unity, respectively.[19]

4 Empirical results

An application of the CSSW model with threshold effects[20], which is the regression-based counterpart of the LCSS model, identified eight groups of banks initially. We did not search for additional groups of banks, as the number of banks in the first three groups was decreasing over time and further segmentation would potentially result in groups with an insufficient number of bank observations to derive reliable conclusions under the LCSS model. In addition, for the conventional stochastic frontier models that rely on order statistics[21], it is common to trim the effects term from the upper and lower 5% percentiles observed at least in one period, to remove the outlier effects.[22] Trimming was not feasible for the smallest two groups of banks as the number of banks in these groups was vanishing towards the end of the sample period. Based on these considerations, we combined the smallest three groups of banks into one group and estimated the LCSS model based on six group-technologies.

Table 2 presents the means and standard deviations of variables in different size groups. Group 1 includes banks with total assets size (K) less than \$1.8 billion (B). Group 2 includes banks with total assets size of more than \$1.8 B, but less than \$2.6 B. Group 3 includes banks with total assets size of more than \$2.6 B, but less than \$5.2 B. Group 4 includes banks with total assets size of more than \$5.2 B, but less than \$16.8 B. Group 5 includes banks with total assets size of more than \$16.8 B, but less than \$53 B. Finally, Group 6 includes banks with total assets size of more than \$53 B.

Table 2 reveals that the average K of Group 1 is slightly less than \$1 B. However, as we mentioned earlier, we dropped the banks with less than \$1 B average asset size from the main estimation sample. The explanation for this seemingly contradictory situation is that some of the banks in Group 1 are smaller banks with K less than \$1 B the first time they appear in the sample, but over time, they grow quickly through mergers or by other means, and they reach larger asset sizes and become a part of other groups with larger banks. Hence, the average asset size of each of these banks over the whole sample time period is more than \$1 B, so they remain in our sample. Nevertheless, the average asset size of each of them while they are in Group 1 is less than \$1 B, and thus, the average K of Group 1 is less than \$1 B.

As we would possibly expect, the means of all other variables are larger in larger bank groups. For example, off-balance sheet items (Y_5) has a mean of \$0.16 B in Group 1, while it has a mean of \$4.16 B in Group 4, and \$84.7 B in Group 6. To give another example, the average number of labor (X_3) in Group 2 is 781, while the same statistic is 1260 in Group 3 and 7878 in Group 5.

What is interesting is that the standard deviations of the variables follow a similar pattern as well that larger bank groups generally have larger standard deviations. This pattern would indicate a larger variation in variables within larger groups, as well as the heterogeneity between different size groups. For instance, commercial and industrial loans (Y_2) has a standard deviation of \$0.12 B in Group 1, while its standard deviation is \$4.11 B in Group 5 and \$26.9 B in Group 6. For the securities (Y_4), the standard deviation is \$0.2 B in Group 1 and \$0.35 B in Group 2, while the same statistic is \$48 B for Group 6. Such differences in standard deviations between groups indicate distributional differences, and such distributional differences foretell different estimation outcomes for different size groups. To sum up, there is variation in the sample even when we only concentrate on the banks with asset size larger than \$1 billion.

Finally, the number of banks in Group 1 is 14,283. This large number is due to aggregating three smallest groups of banks into one group as explained. Excluding Group 1, the number of banks is larger for larger bank groups for Groups 2, 3, and 4, and the number of banks is smaller for the largest bank groups for Groups 5 and 6. There are 4,895 banks in Group 2, while there are 6518 banks in Group 3, and 6867 banks in Group 4. Group 5, on the other hand, has only 2348 banks and Group 6 has 1263 banks. There can be many factors that result in smaller number of banks in the two largest groups compared to the number of banks in Group 4 and other groups. Potentially, one of these factors is a strategy to keep the bank assets below the threshold levels set by banking authorities in order to avoid regulations.

Since communicating the estimation results from the translog functional form is complex, we calculate the means of group-specific marginal effects. In Table 3, we report

[19] See Färe and Primont (2012) for more discussion on the scale elasticity measures for distance functions.
[20] Appendix A explains the CSSW and LCSS models in detail.
[21] For example, the fixed effects estimator of Schmidt and Sickles (1984), CSS estimators, and the Kalman filter estimator of DKS.
[22] See Berger (1993), Berger and Hannan (1998), and Kutlu (2012) for more details.

Table 2 Summary statistics of different size groups

	Group 1 K < \$1.8 B	Group 2 \$1.8 B < K K < \$2.6 B	Group 3 \$2.6 B < K K < \$5.2 B	Group 4 \$5.2 B < K K < \$16.8 B	Group 5 \$16.8 B < K K < \$53 B	Group 6 \$53 B < K
K	\$0.97 B	\$2.18 B	\$3.72 B	\$9.07 B	\$29.60 B	\$196.00 B
	[\$0.48 B]	[\$0.24 B]	[\$0.74 B]	[\$3.00 B]	[\$10.20 B]	[\$234.00 B]
Y_1	\$0.30 B	\$0.74 B	\$1.14 B	\$2.48 B	\$7.70 B	\$45.90 B
	[\$0.23 B]	[\$0.39 B]	[\$0.61 B]	[\$1.47 B]	[\$6.02 B]	[\$60.70 B]
Y_2	\$0.15 B	\$0.30 B	\$0.58 B	\$1.59 B	\$5.66 B	\$29.90 B
	[\$0.12 B]	[\$0.19 B]	[\$0.39 B]	[\$1.01 B]	[\$4.11 B]	[\$26.90 B]
Y_3	\$0.11 B	\$0.23 B	\$0.41 B	\$1.06 B	\$2.62 B	\$14.30 B
	[\$0.12 B]	[\$0.19 B]	[\$0.34 B]	[\$1.19 B]	[\$3.29 B]	[\$19.90 B]
Y_4	\$0.20 B	\$0.38 B	\$0.67 B	\$1.56 B	\$4.32 B	\$27.80 B
	[\$0.20 B]	[\$0.35 B]	[\$0.58 B]	[\$1.37 B]	[\$4.71 B]	[\$48.00 B]
Y_5	\$0.16 B	\$0.53 B	\$0.84 B	\$4.16 B	\$17.40 B	\$84.70 B
	[\$0.63 B]	[\$2.69 B]	[\$1.87 B]	[\$42.60 B]	[\$141.00 B]	[\$326.00 B]
X_1	\$0.14 B	\$0.29 B	\$0.52 B	\$1.25 B	\$3.28 B	\$13.30 B
	[\$0.10 B]	[\$0.17 B]	[\$0.32 B]	[\$0.75 B]	[\$1.85 B]	[\$12.40 B]
X_2	\$0.65 B	\$1.44 B	\$2.36 B	\$5.37 B	\$16.60 B	\$115.00 B
	[\$0.33 B]	[\$0.27 B]	[\$0.58 B]	[\$1.93 B]	[\$6.97 B]	[\$137.00 B]
X_3	369	781	1,260	2,921	7,878	42,111
	[227]	[359]	[613]	[1,454]	[3,677]	[42,054]
X_4	\$0.02 B	\$0.03 B	\$0.05 B	\$0.13 B	\$0.36 B	\$1.85 B
	[\$0.01 B]	[\$0.02 B]	[\$0.03 B]	[\$0.08 B]	[\$0.19 B]	[\$1.63 B]
X_5	\$0.19 B	\$0.45 B	\$0.79 B	\$2.31 B	\$6.98 B	\$24.60 B
	[\$0.18 B]	[\$0.29 B]	[\$0.51 B]	[\$1.72 B]	[\$4.86 B]	[\$22.00 B]
N	14,283	4,895	6,518	6,867	2,348	1,263

B stands for billions. Average values of the variables within groups are presented. Standard deviations are in brackets. In the main text, we explain why the average K of Group 1 is less than \$1 billion

these marginal effects for each size group estimated under the LCSS model.[23] Also, in Appendix D, we provide the full set of thresholds, structural, and distributional parameters, and their standard errors from these regressions.

In Table 3, in the overall, the distributional parameters σ_ε and σ_e are statistically significant at 0.1% significance level, and most of the structural coefficients are statistically significant at 0.1 to 5% significance levels. Since the dependent variable is the negative of natural logarithm of real estate loans ($-\ln(Y_1)$) we would expect the marginal effects of all other output (y) variables to be positive as they are alternatives for y_1. Also, since the dependent variable is $-\ln(Y_1)$, we would expect the marginal effects of all input (x) variables to be negative as they are resources for output variables. Almost all of the mean marginal effects conform with our theoretic expectations. We find that some of the x variable coefficients are positive, but they are not statistically significant so those findings do not contradict the theoretical expectations and considered to be only incidental.

Here, it is important to note that understanding the underlying intuition of the relationships between independent variables and the numeraire output variable (real estate loans) within a regression column is perhaps less important in this paper than observing the differences between different regression results. For example, while the effect of miscellaneous personal loans and credit cards (y_3) is significant and positive in most groups within a relatively small range of 0.105–0.162, the marginal effect of y_3 is smaller and significant in Group 5, and substantially smaller and not significant in Group 6. The insignificant effect of y_3 on y_1 can be hypothesized as how real estate loans would be independent of miscellaneous personal loans and credit cards for the largest banks, but for our analysis in this paper, we are more interested in how the effect of y_3 (or another independent variable) is different for larger size groups than the rest of the size groups as well as the Single-Group.

To give some other examples, the coefficient of off-balance sheet items (y_5) is significant and between 0.009 and 0.023 for Groups 1 to 4. For Groups 5 and 6, however,

[23] To preserve the space, we do not report the estimates of the CSSW model, which are qualitatively similar to the estimates obtained under the LCSS model. The CSSW model estimates for the six groups and the single group are available upon request.

the coefficient is not significant. The effect of capital (x_4) is significant and negative in Groups 1 to 4. While the same effect is also negative in Group 5 and Group 6, it is not significant.

There are also differences between smaller size groups. The effect of total time and savings deposits (x_2) is negative in Groups 1 to 3, but positive and not significant in Group 4. The effect of labor (x_3) is substantial in Group 4, but not so much for the banks in Groups 1. The same coefficient is negative but not significant in Group 3. For Group 2, the coefficient is positive perhaps due to a crowding out effect, but it is not significant. Similarly, the coefficient of x_3 is positive in Groups 5 and 6 but they are not significant. It is important to note once again that our main intention with these examples here is to highlight the differences between groups. Hence, we do not get into details about the underlying intuition of why a coefficient is larger or smaller in one bank group compared to another.

Looking at the findings in Table 3, the differences between columns support that the six size groups identify different bank categories, and they should be assessed based on separate analyses. The size group that is most similar to the Single-Group in terms of coefficient signs, magnitudes, and significance is Group 1. On the other hand, Groups 5 and 6 with the largest banks, seem as the most different groups compared to the Single-Group. Therefore, regulation policies based on the results from the Single-Group would probably address the challenges faced by the smallest banks in Group 1, and such policies may be completely irrelevant for the largest banks in Groups 5 and 6. In order to understand how this situation may become a problem, we look at the bank efficiencies in different groups.

In Figs. 1–6, we depict the average group efficiency scores estimated under the LCSS and CSSW models. In all of these figures, the efficiency trends and rankings estimated under the CSSW model are fairly consistent with those obtained under the LCSS model. While discussing the figures, we mostly concentrate on the estimates under the LCSS model. Figure 1 shows that the group of the smallest banks in our estimation sample, that is the banks with average total assets not exceeding \$1.8 billion (Group 1), has estimated efficiency levels averaged at around 76.3%. This level of efficiency is the lowest compared to that of other size groups. Competitive pressures from new entrants and the inability of these small banks to expand in the highly specialized investment and securities markets could potentially explain the relatively low performance of these banks.

Table 3 Mean marginal effects under LCSS model

	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Single group
y_2	0.202***	0.209***	0.255***	0.306***	0.391***	0.469***	0.251***
	(0.0055)	(0.0076)	(0.0059)	(0.0067)	(0.0177)	(0.0307)	(0.0041)
y_3	0.162***	0.127***	0.105***	0.113***	0.089***	0.033	0.135***
	(0.0051)	(0.0061)	(0.0061)	(0.0073)	(0.0149)	(0.0261)	(0.0043)
y_4	0.032***	0.013***	0.016***	0.020***	0.023***	0.048***	0.023***
	(0.0019)	(0.0021)	(0.0021)	(0.0019)	(0.0040)	(0.0084)	(0.0012)
y_5	0.013**	0.009*	0.017**	0.023***	0.015	0.014	0.016***
	(0.0041)	(0.0042)	(0.0057)	(0.0052)	(0.0106)	(0.0135)	(0.0023)
x_1	−0.127***	−0.621***	−0.569***	−0.587***	−0.224	0.011	−0.389***
	(0.0371)	(0.0668)	(0.0525)	(0.0532)	(0.1533)	(0.1615)	(0.0267)
x_2	−2.338***	−0.212	−0.682***	0.221	0.019	−0.238	−1.189***
	(0.1181)	(0.3835)	(0.1371)	(0.3069)	(0.4754)	(2.5988)	(0.0863)
x_3	−0.367***	0.152	−0.141	−0.887***	0.505	0.837	−0.574***
	(0.0605)	(0.0780)	(0.1479)	(0.0638)	(0.4184)	(1.1759)	(0.0323)
x_4	−0.441***	−0.242**	−0.722***	−1.661***	−0.437	−1.095	−0.733***
	(0.0470)	(0.0896)	(0.1112)	(0.1345)	(0.2805)	(1.1658)	(0.0443)
x_5	−0.718***	−0.500***	−0.368***	−1.070***	−0.499***	−1.116***	−0.834***
	(0.0325)	(0.0698)	(0.0613)	(0.0659)	(0.0908)	(0.2259)	(0.0249)
σ_ε	0.045***	0.037***	0.041***	0.038***	0.048***	0.042***	0.045***
	(0.0005)	(0.0006)	(0.0006)	(0.0005)	(0.0010)	(0.0012)	(0.0002)
σ_e	0.006***	0.007***	0.008***	0.007***	0.007***	0.006***	0.007***
	(0.0002)	(0.0004)	(0.0004)	(0.0003)	(0.0005)	(0.0006)	(0.0001)

y_1 is the numeraire output variable in the equations. Marginal effects of the variables evaluated at the means of the variables are presented. Standard errors are in parentheses. Asterisks indicate significance at the 0.1% (***), 1% (**) and 5% (*) levels

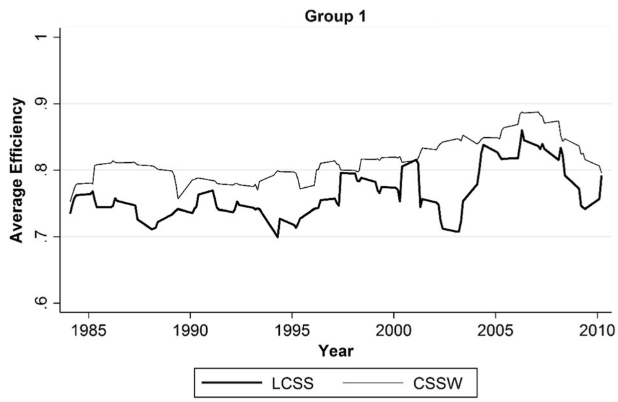

Fig. 1 Group 1 average efficiencies under LCSS vs CSSW models

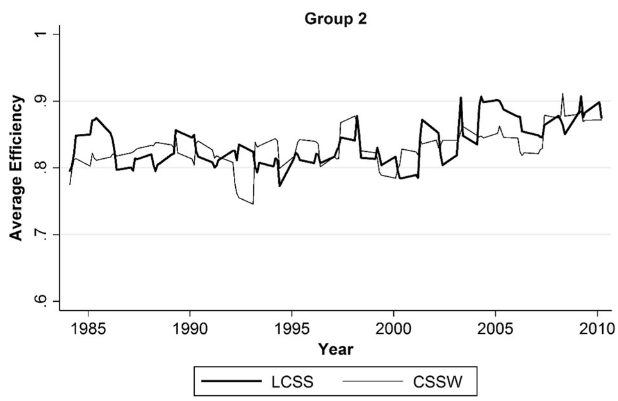

Fig. 2 Group 2 average efficiencies under LCSS vs CSSW models

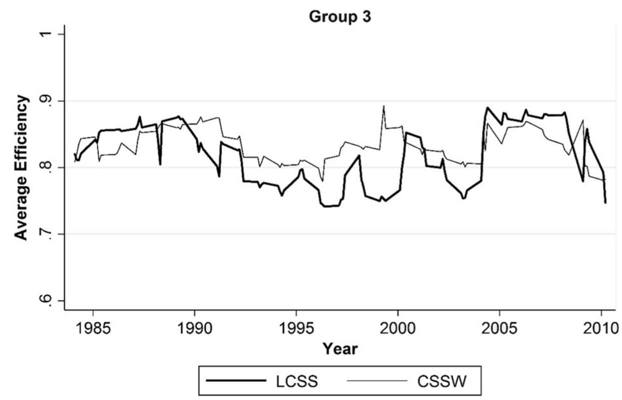

Fig. 3 Group 3 average efficiencies under LCSS vs CSSW models

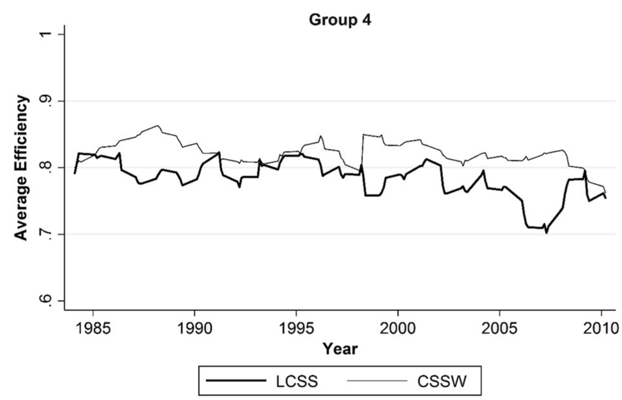

Fig. 4 Group 4 average efficiencies under LCSS vs CSSW models

The estimated average efficiency levels of the mid-sized banks, that is the banks with average total assets ranging from \$1.8 billion to \$16.8 billion (Groups 2 to 4), show a relatively stable trend over time. Their average efficiencies over time are 83.6%, 81.9%, and 78.4%. As presented in Figs. 2–4, on average, Group 4 appears to be a little less efficient than Groups 2 and 3. However, during the period between the second quarter of 2004 and the last quarter of 2009, Group 4 substantially underperformed compared to Group 2 and Group 3 (most of that time, the average efficiency of Group 4 is more than 10% below the average efficiencies of Group 2 and Group 3). This gap in average efficiencies during this period peaks in 2007 to 17.8%.

While the average efficiencies of Group 2 and Group 3 may seem as if they are practically the same, our methodology identifies them as separate groups so there should be technological differences (different frontiers) between these two groups. Over the sample period, the mean difference between the average efficiencies of Group 2 and Group 3 is around 1.7 percentage points. The standard deviation of this difference, however, is around 5 percentage points with the maximum differences at -15.1 (when Group 2 is better performing) and 6.9 (when Group 3 is better performing). Group 2 has a relatively less volatile average efficiency figure that stay within the 80–90% band with a slow but increasing efficiency trend over the sample period. Group 3, on the other hand, has a relatively more volatile average efficiency figure that drops below 80% in 1992 and stays mostly there until 2000. This decline in their efficiencies is probably due to the savings and loan crisis starting in the late 1980s going through the 1990s, and then a subsequent competitive pressure from new entrants due the Riegle-Neal Interstate Banking and Branching Efficiency Act of 1994.

Note that, particularly for that time period between 2004 and 2008, the gap between CSSW and LCSS efficiency estimators in Group 4 is relatively high. CSSW estimator is less flexible and may not be capturing the bank specific efficiency changes as good as the LCSS estimator. From a technical point of view, this may indicate that while the average efficiency is lower for LCSS over that time period, the reason may be an increased level of heterogeneity within the same bank group and the failure of the portfolios of some of them due to specific low performing product categories. Many midsized banks of Group 4 pushed into commercial and construction lending during that time period with the idea that their relatively large size combined

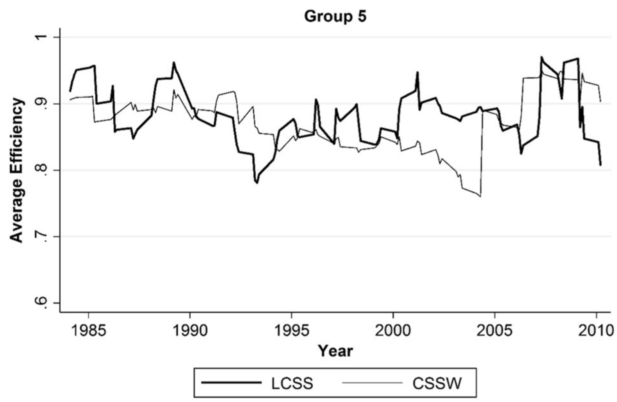

Fig. 5 Group 5 average efficiencies under LCSS vs CSSW models

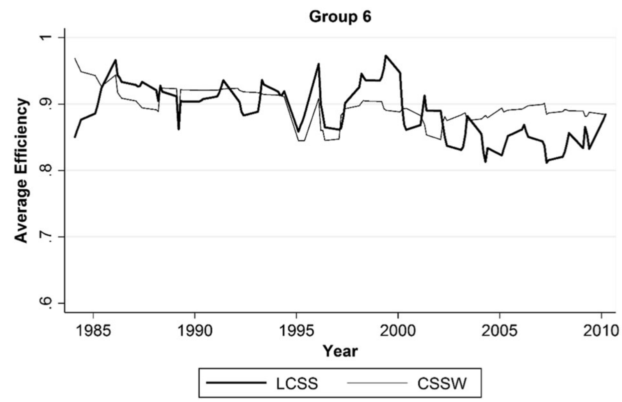

Fig. 6 Group 6 average efficiencies under LCSS vs CSSW models

with the local knowledge would give them a competitive edge in a market dominated by larger banks. On the other hand, some other banks in the same group took a more cautionary approach towards those products. This situation provides a potential explanation of why efficiencies of the banks within Group 4 have a somewhat more heterogeneous efficiency pattern during the years that preceded the recent financial crisis.

To give some examples from that period, Corus Bank, N. A. in Illinois was a Group 4 bank, and their product portfolio was heavily composed of real estate loans, construction loans, and commercial loans. Their efficiency scores were around 44% during the period before the crisis, and in September 2009, the Office of the Comptroller of the Currency closed the bank. MB Financial Bank in Illinois was also a Group 4 bank, but thanks to their relatively more careful approach towards commercial and real estate loans, their efficiency scores were much better at that period varying around 72%. Banks in Groups 2 and 3 perhaps did not try as hard as the banks in Group 4 to compete and catch up with the largest banks and they rather focused on their current portfolio structures, which probably made them more efficient on average in that period compared to Group 4 banks.

The estimated average efficiency scores of Group 5, which is the second largest group of banks with average total assets ranging between \$16.8 B and \$53 B, is estimated to be at 88.3%. Group 6, consisting of the largest banks with average total assets exceeding \$53 B, has the highest efficiency scores on average. The average efficiency level of Group 6 is approximately 88.8% over the sample period. The average efficiency scores of Group 5 and Group 6 tend to be very close to each other. While the mean difference between the average efficiency scores of Group 5 and Group 6 is around 0.4 percentage points, the difference has a 6.5 percentage point standard deviation with the maximum differences at -15.9 (when Group 5 is better performing) and 15.5 (when Group 6 is better performing). The high variation in the difference between the average efficiency scores of Group 5 and Group 6 is an indication that banks in these two groups are not practically the same, and they may respond to different financial settings in different ways.

For instance, as presented in Figs. 5 and 6, Group 5 seems to be somewhat outperforming Group 6 during the period between the third quarter of 2000 and the last quarter of 2009. The mean difference between the average efficiencies of these two groups during this period is around 5 percentage points and the difference peaks to more than 14 percentage points towards the end of 2007. The lower efficiencies of Group 6 relative to the efficiencies of Group 5 over this period can be attributed to the complex structure and "too-big-to-fail" nature of the banks in Group 6. A large number of subsidiaries, branches and offices located nationwide, as well as multi-level management system, makes these banks to be less controllable than those with simpler structures. In addition, the multi-present position and increasing market shares of the banks in Group 6 could possibly allow them to exercise market power, and consequently, this exercise may have a negative effect on their efficiencies.

While not all bank groups responded similarly to the Riegle-Neal Interstate Banking and Branching Efficiency Act of 1994, there was a subsequent boost within the group-specific performance rates of some banks as a result of the enactment. The nationwide banking and branching particularly benefited the large banks in Group 5 and Group 6, which were able to expand their operations beyond the state borders. Especially for the banks in Group 5, there is an observable increase in average efficiencies starting in 1994, and the effect lasts longer perhaps through 2001. For some others banks, though, the efficiency boost of the act seems only as temporary. For example, the banks in Group 6 experience a sharp increase in average efficiency in 1994 with a sharp drop in 1996.

There is also a second wave of augmented performance, in particular for the largest banks in Group 6, due to the securitization of illiquid assets, such as loans, that was permitted under the Gramm–Leach–Bliley Act of 1999. Our efficiency estimates suggest that industrial and digital/electronic improvements and an introduction of new financial products in the early 2000s, along with the U.S. housing market boom of 2004–2005, may have helped banks in different groups to further enhance their performance levels, and for certain groups of banks, to reach considerably high efficiency levels. For example, Group 3 banks experience an increase in the average efficiencies by more than 9% in the early 2000s, and another increase by more than 12.5% between 2003 and 2005. Westamerica Bank in California is a Group 3 bank in the 1990s and 2000s, and they started providing online services, and niche products and services such as VIP Banking for their individual customers in late 90s.[24] Our results show that their efficiency during that period increased from 70.4 to 91%. Later with the real estate loans supported by core deposits, they strengthen their performance through 2000s. While their efficiency in the third quarter of 2003 is at 79.8% towards the end of 2005, their efficiency level becomes as high as 98.7%.

Many banks experienced a decline in their performance during the recent financial crisis of 2007–2010. We observe that some of these banks responded to this shock slower than others and some other banks, such as the ones in Groups 2 did relatively better on average compared to other groups during the crisis. The banks in Group 6 showed an immediate decrease in their efficiencies due to the peak of the delinquency rate in residential mortgage backed securities resulting from the collapse of the U.S. housing bubble in the summer of 2006.

Lastly, the banks in all six groups appear to have experienced diseconomies of scale over the sample period. The median values of returns to scale estimates for the bank groups range between 0.78 and 0.96, which indicate that the banks operate above the optimal output level and, ceteris paribus, can improve their scale efficiencies by reducing their output levels and achieve optimal input combinations. Nevertheless, the overall median value is 0.87, which is not significantly different from constant returns to scale at 5% significance level.[25] Previous studies of the U.S. commercial banking industry have shown that, on average, banks operated at constant returns to scale within the period under consideration.[26]

To compare and contrast the efficiency scores and rankings from the different size groups model to those

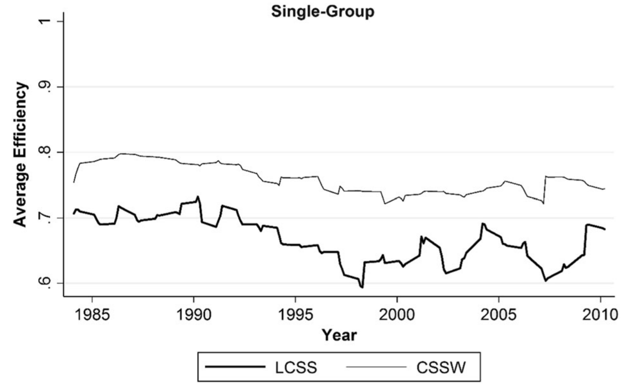

Fig. 7 Aggregated Single-Group average efficiencies under LCSS and CSSW

obtained with the single aggregated group model, we also plot the efficiency estimates obtained under the assumption of a Single-Group for both LCSS and CSSW models in Fig. 7. It is evident from Fig. 7 that the estimated common technology/frontier efficiency scores are considerably lower than the efficiency scores estimated under the heterogeneous technologies assumption. The average efficiency of the Single-Group over the sample time period is around 66.7% which is closer to the average efficiency of Group 1 banks. This finding has interesting implications. First, the largest banks may seem as if they are less efficient under the Single-Group common frontier assumption. For example, the banks in Group 6 have an average efficiency score at around 88.8% according to the heterogeneous group estimates. So, in the aggregated analysis of bank efficiencies, they appear as 22% less efficient than how they are in the heterogeneous groups analysis.

Secondly, under the Single-Group common frontier assumption, smaller banks may appear as more efficient than they are relative to their larger counterparts. For instance, according to the heterogeneous frontiers model, the average efficiency of Group 1 is 76.3%, while the average efficiency of Group 5 is 88.3%. However, according to the common frontier model, all of the banks in Group 1 and Group 5 are at 66.7% average efficiency over time. That is, even though the banks in Group 1 are less efficient on average compared to the banks in Group 5 according to the heterogeneous frontiers model, they all appear to be equally efficient on average in single frontier model. This would imply that in the single frontier model, some of the Group 1 banks appear as relatively more efficient than their counterparts in Group 5. To give an example, according to the heterogeneous frontiers model, in the first quarter of 1992, InterWest Savings Bank in Oak Harbor, Washington is a Group 1 bank with 76.96% efficiency, while CoreStates National Bank in Philadelphia, Pennsylvania is a Group 5 bank with 85.03% efficiency. According to the single frontier model, however, the efficiency of the InterWest

[24] The details are provided in the website of the bank at www.westamerica.com

[25] This is based on the empirical distribution of the returns to scale values around their median values.

[26] See for example McAllister and McManus (1993) and Wheelock and Wilson (2001).

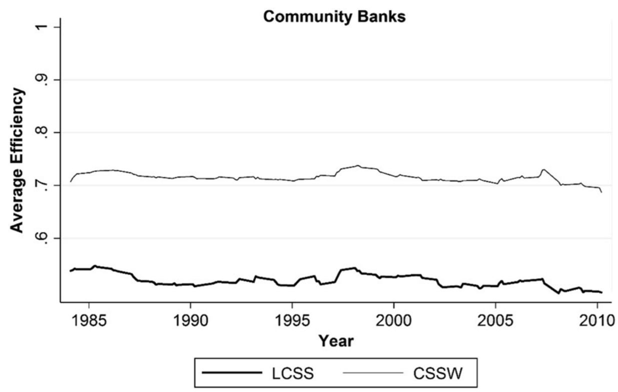

Fig. 8 Average efficiencies of community banks under LCSS vs CSSW models

Savings Bank in the first quarter of 1992 is 92.19%, while the efficiency of the CoreStates National Bank is 77.81%. So, while InterWest is less efficient than CoreStates according to the heterogeneous frontiers model, InterWest seems substantially more efficient than CoreStates in the single frontier model.

Finally, with Fig. 7, we find evidence that the common frontier model cannot capture important historical events and periods as well as the heterogeneous frontiers model. During the period between 1985 and 1994, the average efficiency of the aggregated Single-Group is around 70% without a sharp drop in efficiencies or a clear indication of the savings and loan crisis. Also, the gains in efficiencies due to the Riegle-Neal Interstate Banking and Branching Efficiency Act of 1994 are not observable in the Single-Group figure. According to Fig. 7, during that period, the average efficiency of the banks is somewhat flat at around 65.9% with a slight downward trend. Subsequent boosts in performance due to the Gramm–Leach–Bliley Act of 1999, new financial products in the early 2000s, and the U.S. housing market boom of 2004-2005, are not well captured in Fig. 7 either. Average efficiency during that time period is well below 70% in Fig. 7, while many banks are achieving efficiencies higher than 80% on average according to Figs. 1–6. Lastly, the Single-Group figure dips to 60% average efficiency during the most recent financial crisis with a following increase in efficiencies. While this part of the figure may be giving a rough idea of what is going on in the economy in general, it does not portray what the bank groups, especially the banks in Groups 3 and 5, experience during that period.

In addition to the Single-Group analysis presented above, we were also able to analyze all banks with less than $1 billion average real asset size over time despite the computational difficulties associated with such analysis. We plot the efficiency estimates of these community banks for both LCSS and CSSW models in Fig. 8. Based on the results we discussed earlier, we would probably expect to see that the community banks are the least efficient banks on average compared to the other banks in Groups 1 to 6. Figure 8 demonstrates that with an average efficiency around 52.6%, it is in fact the case that community banks are less efficient than the banks in Group 1. Also, the average efficiency of the community banks is substantially below the aggregated Single-Group efficiencies illustrated in Fig. 7, and they show different trend patterns. For example, while the Single-Group efficiency figure shows a falling trend from 1992 to 1998 with a sharp drop to 59.4% in the third quarter of 1998, a similar trend does not seem to exist for community banks during the same period. To give another example, after 2007, the Single-Group average efficiency exhibits a steady upturn while the average efficiency of community banks plummets during that period.

It is clear that the Single-Group efficiency figure does not represent the average efficiencies of bank groups presented in other figures. Evidently, there are considerable efficiency differences between different groups that our methodology identifies, and they are all different than what a single frontier model would indicate. These implications may have serious consequences. Suppose that a team of economists examine the efficiencies of these banks with the aggregated data and the assumption of a common frontier, and find that the banks are inefficient in overall with the largest banks at extreme inefficiencies. Looking at these findings, policy advisors can come up with banking regulations that target the inefficiencies of the banks. Since the efficiency scores of Group 1 is the closest to the Single-Group efficiency scores, these policies may become somehow relevant to and useful for Group 1 banks, and perhaps for the smallest community banks. However, the Single-Group assumption does not result in accurate estimates of efficiencies for the largest banks, so the policies based on Single-Group findings may turn out to be completely irrelevant to the banks in Groups 5 and 6. In the worst scenario, these banking regulations may be too restrictive to an extent that diminishes the efficiencies of the largest banks and self-fulfills the Single-Group efficiency findings. The potential outcomes of these regulations may yield questions such as if the banking authorities are looking at the wrong end of the size spectrum.

Compliance with the type of banking regulations outlined above would be costly for banks, and their expenses would increase as the regulations become stricter. Tighter regulations on the largest banks may result in burdens, such as additional inefficiencies that are not proportional to the benefits that these regulations provide to the economy. Also, there may be some unintended consequences of requiring the largest banks to hold more capital. For example, the purpose of a threshold-based risk-weighted asset requirement is to reduce the probability of a default and reduce the severity of outcomes in case of a default. The same requirement, however, may create incentives for the large banks to stay below the threshold levels to avoid regulations, and some other incentives for the largest banks that cannot escape the regulations to innovate products, which in turn, may add to the buildup of a financial crisis. Slovik (2012) explores this possible connection between the strict capital regulations based on risk-weighted assets and financial crises.

Suppose that instead of focusing on the banks as a Single-Group, this time the team of economists only concentrates on the largest banks. Suppose further that the team arbitrarily establishes an asset size as a threshold to determine the systemic financial importance of these banks and sort them into two groups. The banks with capital above the threshold level will be subject to enhanced prudential standards including higher capital requirements to account for their significance. If the requirement to hold more capital results in extra inefficiencies, this situation would create an incentive for the banks on the borderline to intentionally hold their assets below the line, perhaps with anticipation before the enactment or enforcement of these regulations, to avoid the inefficiencies that would come with the importance status. The Basel Accords issued by the Basel Committee on Banking Supervision, and the Dodd-Frank Wall Street Reform and Consumer Protection Act of 2010 signed by President Obama introduce such regulations with threshold levels that are not necessarily based on apparent analytical foundations, but are used to determine and focus on the Systemically Important Financial Institutions (SIFI). The implementations of Basel III and Dodd-Frank Act are beyond the data sample of our study, but we would probably expect to see similar banking strategies for efficiency incentives under the minimum capital requirements of Basel II that address risks, and during the phase when Basel III was negotiated.

Looking at Figs. 5 and 6, what we observe between 2007 and 2010 may be partially an artifact of this phenomenon. Many banks in Group 5 outperform the banks in Group 6 during this period. To give some examples, Compass Bank had an asset size of \$46.8 B on average between 2007 and 2010, and as a Group 5 bank, the bank's average efficiency over this period is estimated to be approximately 87.1%. Capital One, on the other hand, had an \$87.4 B asset size on average over the same period, and the bank has been designated as a Domestic Systemically Important Bank (D-SIB) since 2009, which makes Capital One subject to the stringent annual Stress Tests of the Federal Reserve. Capital One is a Group 6 bank in our study and has a 76.15% average efficiency between 2007 and 2010. JPMorgan Chase is one of the biggest banks in Group 6, and the bank is identified as one of the few U.S. bank holding companies that are global systemically important banking organization (G-SIB). JPMorgan Chase's average efficiency is more than 95.9% over the 1990s, but between 2007 and 2010, the efficiency drops below 75% with an average around 68.7% and a minimum at 65%, which is even below the efficiencies of many Group 1 banks during the same period. With these findings, we cannot help raise questions about if these efficiency differences between Group 5 and Group 6 banks over the 2007–2010 period are completely due to the financial crisis experienced to some extent by all banks or if regulations that specifically target the banks in Group 6 play an important role in lowering their efficiencies while giving incentives to Group 5 banks and others to perform better than Group 6 banks.

5 Conclusions

In this paper, we applied our model with threshold effects to study the performance of the U.S. commercial banks during the 1984–2010 period. By employing the asset size of banks as a threshold variable, we identified the existence of six size groups and estimated time-varying bank-specific and group-specific inefficiency scores, and scale elasticity estimates. We found that, on average, the two top-tier size groups of banks were operated more efficiently than the smaller size groups of banks over the sample period. The estimated efficiency trends for each group were consistent with and largely explained by the effects of deregulatory reforms in the U.S. banking industry, banking crises, as well as by the effects of financial, industrial and digital innovations. We showed that estimating the model under the assumption of a common frontier results in considerably lower average efficiency scores which cannot capture the important historical events and periods as well as the average group efficiency scores from the heterogeneous frontiers model. Also, under the common frontier assumption, the largest banks appear to be 22% less efficient on average than how they are according to our heterogeneous frontiers model. Furthermore, in the common frontier model, smaller banks appear to be relatively more efficient on average compared to their larger counterparts. This would mean that policies or regulations based on common frontier measures may be stricter or looser than necessary, which in turn, may result in unintended consequences. Finally, we found that the banks in our sample have been experiencing diseconomies of scale over the sample period, that is, they have been operating at or above the optimal scale level. Although the point estimates indicate that all the bank groups are subject to decreasing returns to scale, we could not reject constant returns to scale for the whole sample.

Compliance with ethical standards

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

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

Adams RM, Berger AN, Sickles RC (1999) Semiparametric approaches to stochastic panel frontiers with applications in the banking industry. J Bus Economic Stat 17:349–358

Ahn SC, Sickles RC (2000) Estimation of long-run inefficiency levels: a dynamic frontier approach. Econom Rev 19:461–492

Aigner DJ, Lovell CAK, Schmidt P (1977) Formulation and estimation of stochastic frontier production function models. J Econ 6:21–37

Alam IMS (2001) A nonparametric approach for assessing productivity dynamics of large US banks. J Money Credit Bank 33:121–139

Almanidis P (2013) Accounting for heterogeneous technologies in the banking industry: a time-varying stochastic frontier model with threshold effects. J Product Anal 39:191–205

Baltensperger E (1980) Alternative approaches to the theory of the banking firm. J Monetary Econ 6:1–37

Battese EG, Heshmati A, Hjalmarsson L (2000) Efficiency of labour use in the swedish banking industry: A stochastic frontier approach. Empir Econ 25:623–640

Battese GE, Corra GS (1977) Estimation of a production frontier model: With application to the pastoral zone of eastern australia. Aust J Agric Econ 21:169–179

Berger AN (1993) "Distribution-free" estimates of efficiency in the US banking industry and tests of the standard distributional assumptions. J Product Anal 4:261–292

Berger AN, Hannan TH (1998) The efficiency cost of market power in the banking industry: a test of the "quiet life" and related hypotheses. Rev Econ Stat 80:454–465

Berger AN, Hasan I, Zhou M (2009) Bank ownership and efficiency in china: What will happen in the world's largest nation? J Bank Financ 33:113–130

Berger AN, Hasan I, Zhou M (2010) The effects of focus versus diversification on bank performance: Evidence from chinese banks. J Bank Financ 34:1417–1435

Berger AN, Humphrey DB (1992) Measurement and efficiency issues in commercial banking, Output measurement in the service sectors. University of Chicago Press, Chicago, IL, USA, pp. 245–300

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

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

Brewer E, Jagtiani J (2013) How much did banks pay to become too-big-to-fail and to become systemically important? J Financial Serv Res 43:1–35

Brissimis SN, Delis MD, Tsionas EG (2010) Technical and allocative efficiency in European banking. Eur J Operational Res 204:153–163

Chang T-P, Hu J-L, Chou RY, Sun L (2012) The sources of bank productivity growth in china during 2002–2009: a disaggregation view. J Bank Financ 36:1997–2006

Christopoulos DK, Tsionas EG (2001) Banking economic efficiency in the deregulation period: results from heteroscedastic stochastic frontier models. Manch Sch 69:656–676

Coelli TJ, Perelman S (1999) A comparison of parametric and non-parametric distance functions: with application to European railways. Eur J operational Res 117:326–339

Coelli TJ, Perelman S (2000) Technical efficiency of European railways: A distance function approach. Appl Econ 32:1967–1976

Cornwell C, Schmidt P, Sickles RC (1990) Production frontiers with cross-sectional and time-series variation in efficiency levels. J Econ 46:185–200

Davies RB (1987) Hypothesis testing when a nuisance parameter is present only under the alternative. Biometrika 74:33–43

Delis M, Iosifidi M, Tsionas MG (2017) Endogenous bank risk and efficiency. Eur J Operational Res 260:376–387

Demirgüç-Kunt A, Huizinga H (2013) Are banks too big to fail or too big to save? International evidence from equity prices and cds spreads. J Bank Finance 37:875–894

DeYoung R (2013) Modeling economies of scale in banking. Efficiency and Productivity Growth: Modelling in the Financial Services Industry. John Wiley & Sons, Ltd, West Sussex, UK, pp. 49–75

Dong Y, Firth M, Hou W, Yang W (2016) Evaluating the performance of chinese commercial banks: a comparative analysis of different types of banks. Eur J Operational Res 252:280–295

Durbin J, Koopman SJ (2012) Time series analysis by state space methods. Oxford University Press, Oxford, UK

Duygun M, Kutlu L, Sickles RC (2016) Measuring productivity and efficiency: a Kalman filter approach. J Product Anal 46:155–167

El-Gamal MA, Inanoglu H (2005) Inefficiency and heterogeneity in Turkish banking: 1990–2000. J Appl Econ 20:641–664

Färe R, Primont D (1990) A distance function approach to multioutput technologies. South Econ J 56:879–891

Färe R, Primont D (2012) Multi-output production and duality: Theory and applications. Springer Science & Business Media, New York, NY, USA

Frank B (2015) Getting Frank on Dodd-Frank. Third Way. http://www.thirdway.org/transcript/getting-frank-on-dodd-frank

Galán JE, Veiga H, Wiper MP (2015) Dynamic effects in inefficiency: evidence from the colombian banking sector. Eur J Operational Res 240:562–571

Greene W (2005) Reconsidering heterogeneity in panel data estimators of the stochastic frontier model. J Econ 126:269–303

Hansen BE (1996) Inference when a nuisance parameter is not identified under the null hypothesis. Econometrica 64:413–430

Hansen BE (1999) Threshold effects in non-dynamic panels: estimation, testing, and inference. J Econ 93:345–368

Hansen BE (2000) Sample splitting and threshold estimation. Econometrica 68:575–603

Huang Y (2004) Appendix b: a guide to state operating aid programs for elementary and secondary education, In: Yinger J (ed) Helping children left behind: state aid and the pursuit of educational equity. MIT Press, Cambridge, MA

Hultberg PT, Nadiri MI, Sickles RC (2004) Cross-country catch-up in the manufacturing sector: Impacts of heterogeneity on convergence and technology adoption. Empir Econ 29:753–768

Isik I (2007) Bank ownership and productivity developments: evidence from turkey. Stud Econ Financ 24:115–139

Kalman RE (1960) A new approach to linear filtering and prediction problems. J Fluids Eng 82:35–45

Kaparakis EI, Miller SM, Noulas AG (1994) Short-run cost inefficiency of commercial banks: a flexible stochastic frontier approach. J Money Credit Bank 26:875–893

Kraft E, Tırtıroğlu D (1998) Bank efficiency in croatia: a stochastic-frontier analysis. J Comp Econ 26:282–300

Kumbhakar SC, Wang D (2007) Economic reforms, efficiency and productivity in chinese banking. J Regulatory Econ 32:105–129

Kutlu L (2012) US banking efficiency, 1984–1995. Econ Lett 117:53–56

Lang G, Welzel P (1999) Mergers among german cooperative banks: a panel-based stochastic frontier analysis. Small Bus Econ 13:273–286

Lin X, Zhang Y (2009) Bank ownership reform and bank performance in china. J Bank Financ 33:20–29

McAllister PH, McManus D (1993) Resolving the scale efficiency puzzle in banking. J Bank Financ 17:389–405

Meeusen W, van den Broeck J (1977) Efficiency estimation from Cobb-Douglas production function with composed errors. Int Economic Rev 18:435–444
Mehdian S, Perry M, Rezvanian R (2007) The effect of globalization on efficiency change, technological progress and the productivity growth of U.S. small and large banks. N Am J Finance Bank Res 1:1–21
Mukherjee K, Ray SC, Miller SM (2001) Productivity growth in large US commercial banks: the initial post-deregulation experience. J Bank Financ 25:913–939
Muñiz M (2002) Separating managerial inefficiency and external conditions in data envelopment analysis. Eur J Operational Res 143:625–643
O'Donnell CJ, Coelli TJ (2005) A bayesian approach to imposing curvature on distance functions. J Econ 126:493–523
Oliveira RdF, Schiozer RF, Barros LAdC (2015) Depositors' perception of "too-big-to-fail". Rev Financ 19:191–227
Orea L, Kumbhakar SC (2004) Efficiency measurement using a latent class stochastic frontier model. Empir Econ 29:169–183
Paradi JC, Zhu H, Edelstein B (2012) Identifying managerial groups in a large canadian bank branch network with a dea approach. Eur J Operational Res 219:178–187
Schmidt P, Sickles RC (1984) Production frontiers and panel data. J Bus Economic Stat 2:367–374
Sealey CW, Lindley JT (1977) Inputs, outputs, and a theory of production and cost at depository financial institutions. J Financ 32:1251–1266
Sickles RC (2005) Panel estimators and the identification of firm-specific efficiency levels in parametric, semiparametric and nonparametric settings. J Econ 126:305–334
Slovik P (2012) Systemically Important Banks and Capital Regulation Challenges. Working Paper No 916. OECD Economics Department
Tsionas EG (2002) Stochastic frontier models with random coefficients. J Appl Econ 17:127–147
Tsionas MG (2017) Microfoundations for stochastic frontiers. Eur J Operational Res 258:1165–1170
Vivas AL (1997) Profit efficiency for Spanish savings banks. Eur J Operational Res 98:381–394
Wheelock DC, Wilson PW (1995) Explaining bank failures: deposit insurance, regulation, and efficiency. Rev Econ Stat 77:689–700
Wheelock DC, Wilson PW (1999) Technical progress, inefficiency, and productivity change in US banking, 1984–1993. J Money Credit Bank 31:212–234
Wheelock DC, Wilson PW (2001) New evidence on returns to scale and product mix among US commercial banks. J Monetary Econ 47:653–674