Empir Econ (2015) 48:143–168
DOI 10.1007/s00181-014-0831-9

Nonparametric estimation of returns to scale using input distance functions: an application to large U.S. banks

Diego Restrepo-Tobón · Subal C. Kumbhakar

Received: 17 June 2013 / Accepted: 12 April 2014 / Published online: 6 June 2014

Abstract We derive new measures of returns to scale based on input distance functions (IDFs) and estimate them using nonparametric regression methods. In contrast to the cost function approach, the IDF does not require input prices which are usually unavailable or measured imprecisely. In addition, we can account for equity and physical capital in the IDF. These variables are either excluded from the analysis (especially in a cost function approach) or treated as quasi-fixed inputs, because their prices are not readily available. In our application, we use data for bank holding companies and large commercial banks in the U.S. from 2000 to 2010. We find that although some of these institutions enjoy increasing returns to scale, scale economies are economically small. Thus, concerns about potential cost increases arising from breaking up large banking organizations seem exaggerated, especially from the scale economies point of view.

Keywords Nonparametric regression · Returns to scale · Distance functions · Banks

JEL Classification D24 · G21 · L13 · C14

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

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, NY13902, USA
e-mail: kkar@binghamton.edu

1 Introduction

More than 3 years after The Dodd–Frank Wall Street Reform and Consumer Protection Act was signed into law, regulators, policymakers, and academics in the U.S., the U.K., and elsewhere are still pondering the possibility and desirability of limiting the size of large banking organizations. Recent speeches by top Federal Reserve Bank officials in the U.S. and of the Bank of England in the U.K. revived the unsettled debate on "too big to fail" (TBTF) and the feasibility of limiting the scale and scope of bank activities, calling for further research on banking industry structure in general and on economies of scale and scope in particular.[1]

If large banking organizations enjoy economies of scale, limiting or decreasing their size may impose substantial losses on the economy and great challenges to regulators. However, recent research focusing on the existence of economies of scale and on the cost of shrinking or capping the size of banks is still inconclusive. Wheelock and Wilson (2011, 2012) present evidence indicating that all U.S. commercial banks, bank holding companies (BHC), and credit unions operate under increasing returns to scale (RTS). Likewise, Hughes and Mester (2013) find evidence of increasing RTS for bank holding companies (BHC) operating in 2007. For large U.S. commercial banks, those with assets in excess of $1 billion, Feng and Serletis (2010) present evidence of slightly increasing RTS, Feng and Zhang (2012) find decreasing RTS, and Restrepo-Tobón et al. (2012) find increasing RTS for about 73 % of the banks and constant or slightly decreasing RTS for the rest.

We contribute to this literature and the policy debate by deriving new measures of RTS based on input distance functions (IDFs) and estimating them nonparametrically. In contrast to a cost function approach, the IDF gives a primal representation of the underlying technology, and its estimation requires no information on cost and input prices. Studies focused on conventional economies of scale based on cost functions can be potentially misleading, since they may identify lower funding costs of TBTF as evidence of scale economies. RTS estimates from the IDF are, however, partially shielded from this problem, because the IDF uses no information on costs or input prices, and therefore, estimates from the IDF are less likely to be affected by funding advantages by the big banks. Additionally, nonparametric methods circumvent the potential misspecification problem inherent in parametric models and dispense with assumptions about the true unknown functional form of the underlying production technology.

Use of the IDF allows us to account for equity and physical capital in the estimation process. Many studies based on cost functions exclude these two important variables from the analysis or treat them as quasi-fixed inputs, because their prices are not readily available [e.g., Wheelock and Wilson (2012) and Berger and Mester (2003)]. Output distance functions (ODFs) also share these advantages [see Feng and Serletis (2010) and Feng and Zhang (2012)]. However, the ODF gives biased and inconsistent estimates when inputs are endogenous, which is the underlying assumption in the

[1] See Tarullo (2012a, b, c), Haldane (2012), Rosenblum (2011), and the ensuing media coverage and market participants' analyses [e.g., Johnson (2012), Wack (2012), Wallison (2012), and Harrison (2012)].

cost-minimizing case. In contrast, the IDF does not suffer from this problem [see Das and Kumbhakar (2012)].

The papers closest to our work are Wheelock and Wilson (2011, 2012) (henceforth W&W). However, our paper differs from theirs in several respects. First, while W&W use a cost function, we use the IDF. Second, we derive the ray-scale economies (RSEs) measure from the axiomatic properties of the underlying bank technology represented by cost, output distance, and input distance functions. We show that the expansion path scale economies (EPSE) measure is the ratio between two RSEs evaluated at two different points. As a result, EPSE may not accurately measure the nature and magnitude of RTS in some cases. Third, we study a more recent time period which is more relevant for current policy debates and consider only U.S. commercial banks and BHC with assets above $500 million. The existence of scale economies at smaller banking organizations is not an issue in the literature.[2] Finally, it is often argued that the gradient-based measure of RTS from nonparametric functions might be noisy. We compute such measures and find that RSE estimates are similar to those derived from gradient-based measures.

In line with conventional wisdom, we find that not all BHC and commercial banks enjoy increasing returns to scale (IRTS). In addition, economies of scale for those banking organizations operating under IRTS are small. Our RTS estimates are generally close to unity. Thus, despite the presence of IRTS for some of the biggest banking organizations, the cost of breaking up some of these institutions into smaller and more manageable organizations may not impose heavy costs on the economy.

In the next section we derive our measures of RTS and compare them with those presented in related studies. Sect. 3 presents our model (production technology for banks) and describes the data. Sect. 4 discusses the econometric estimation of the model, and Sect. 5 presents the main empirical results. We conclude by highlighting the policy implications in Sect. 6.

2 Methodology

In this section we derive our RTS measures for the banks based on the axiomatic properties of the underlying production technology. We use cost and input distance functions and highlight the main differences with related measures appearing in the literature.[3]

2.1 Modeling banks' technology

We assume that banks have a production technology $\mathbb{T} : \mathbb{R}_+^N \times \mathbb{R}_+^M \to T = \{(x, y) : x \text{ can produce } y\}$ that transforms input vectors $x = (x_1, ..., x_N) \in \mathbb{R}_+^N$ into output

[2] To avoid the problem that differences in regulations might contaminate the estimates of RTS, we study the period after deregulation of the U.S. banking industry from 2001 to 2010. This strategy allows us to have a smaller sample and to avoid the use of the principal component analysis to mitigate the dimensionality problem highlighted in W&W.

[3] The derivation of our proposed measures of RTS using ODFs is analogous.

vectors $y = (y_1, ..., y_M) \in \mathbb{R}_+^M$. The output correspondence $\mathbb{P} : \mathbb{R}_+^N \to P(x) = \{y : (x, y) \in T\}$ maps the input vectors (x) into output sets $P(x)$ which contain all producible output vectors from $x \in \mathbb{R}_+^N$. The input correspondence $\mathbb{L} : \mathbb{R}_+^M \to L(y) = \{x : (x, y) \in T\}$ maps output vectors y into input sets $L(y)$ which contain all input vectors that can produce $y \in \mathbb{R}_+^M$. By definition, $y \in P(x) \iff (x, y) \in T \iff x \in L(y)$. Thus, $\mathbb{T}$, $\mathbb{P}$, and $\mathbb{L}$ are equivalent representations of the production technology.

The cost and input distance functions are defined, respectively, as follows:[4]

$$C(y, w) = \min_x \{px : x \in L(y)\}, \quad y \in \text{Domain}L(y), \quad w > 0. \tag{1}$$

$$D_{\text{I}}(y, x) = \sup_\lambda \{\lambda > 0 : (x/\lambda) \in L(y)\} \quad \text{for all } y \in \mathbb{R}_+^M. \tag{2}$$

Under minimal standard assumptions, the cost function is a dual representation of the underlying technology. Moreover, the IDF is a primal representation of the production function. The properties of the cost function are well known, and those of the IDFs are detailed in Färe (1988). For instance, the IDF is nondecreasing and linear homogeneous in x and nonincreasing and, in general, not homogeneous in y. We exploit these properties to derive our IDF-based measures of RTS.

Following (see Färe 1988) it is easy to show that the production technology exhibits nonincreasing returns to scale (NIRTS) if and only if for all $\gamma \geq 1$ the following (subhomogeneity) conditions hold:

$$\text{I.} \quad P(\gamma x) \subseteq \gamma P(x)$$
$$\text{II.} \quad L(\gamma y) \subseteq \gamma L(y)$$

For $\lambda \in (0, 1]$ the signs in I and II are reversed. If the opposite (superhomogeneity) conditions hold, then the technology exhibits nondecreasing returns to scale (NDRTS). In addition, the technology exhibits constant returns to scale (CRTS) if the signs in I and II are changed to equality (homogeneity conditions).

Under NIRTS and for $\gamma \geq 1$, conditions I and II imply the following relations for the cost and IDF:

$$\gamma\, C(y, w) \leq C(\gamma\, y, w) \tag{3}$$

$$(1/\gamma) D_{\text{I}}(y, x) \geq D_{\text{I}}(\gamma\, y, x) \tag{4}$$

For $\gamma \in (0, 1)$ the signs in (3) and (4) have to be reversed. The conditions for NDRTS are obtained by reversing the inequality signs. Likewise, the conditions for CRTS are obtained by substituting the inequality signs to equality signs.[5]

[4] See Färe and Primont (1995).

[5] Similar definitions can be found for ODFs.

We prove (4) for $\gamma \geq 1$. From (2)

$$\begin{aligned} D_{\mathrm{I}}(\gamma\, y, x) &= \sup_{\lambda}\left\{\lambda > 0 : \frac{x}{\lambda} \in L(\gamma\, y)\right\} \leq \sup_{\lambda}\left\{\lambda > 0 : \frac{x}{\lambda} \in \gamma\, L(y)\right\} \\ &= \sup_{\lambda}\left\{\lambda > 0 : \frac{x}{\lambda\gamma} \in L(y)\right\} = \frac{1}{\gamma}\sup_{\lambda}\left\{\lambda\gamma > 0 : \frac{x}{\lambda\gamma} \in L(y)\right\} \\ &= \frac{1}{\gamma}\sup_{\phi}\left\{\phi > 0 : \frac{x}{\phi} \in L(y)\right\} = \frac{1}{\gamma} D_{\mathrm{I}}(y, x) \end{aligned} \tag{5}$$

where the inequality follows from condition II. Condition (3) can be proved analogously.

According to condition (3), under NIRTS if outputs were increased by a factor $\gamma > 1$, total costs would increase by a factor greater than γ. Otherwise, with fixed output prices, it would be profitable to expand production, since revenues would increase by a factor γ, but total costs would increase by a factor less than γ. The economic intuition of condition (4), on the other hand, is less evident and requires familiarity with the concept of the input distance function—we return to this issue in the next subsection. For this reason, only some versions of (3) have been used in the literature to measure the nature of scale economies using discrete changes in outputs and inputs.

2.2 Measuring returns to scale

Eliminating the inequalities in conditions (3)–(4), we can define the following measures of RTS based on the cost and input distance functions as follows.

RTS based on the cost function:

$$S^{\mathrm{cost}}(\gamma|y, w) = \frac{C(\gamma\, y, w)}{\gamma C(y, w)} \tag{6}$$

For $\gamma > 1$, the technology exhibits NIRTS, CRTS, or NDRTS if $S^{\mathrm{cost}}(\gamma|y, w) \gtreqless 1$, respectively. For $\gamma \in (0, 1]$, the signs need to be reversed.

RTS based on the IDF:

$$S^{\mathrm{IDF}}(\gamma|y, x) = \frac{D_{\mathrm{I}}(\gamma\, y, x)}{(1/\gamma) D_{\mathrm{I}}(y, x)} \tag{7}$$

For $\gamma > 1$, the technology exhibits NIRTS, CRTS, or NDRTS, if $S^{\mathrm{IDF}}(\gamma|y, x) \lesseqgtr 1$, respectively. For $\gamma \in (0, 1]$, the signs need to be reversed.

Since the IDF is linear homogeneous in x, (7) can be rewritten as

$$S^{\mathrm{IDF}}(\gamma|y, x) = \frac{D_{\mathrm{I}}(\gamma\, y, \gamma\, x)}{D_{\mathrm{I}}(y, x)} \tag{8}$$

which makes it clear that $S^{\mathrm{IDF}}(\gamma|y, x)$ measures how the distance function changes when all outputs and inputs are scaled up or down by the scaling factor γ. If the IDF

does not change, it means that the technology exhibits CRTS: an equiproportional increase in all inputs leads to an equal equiproportional increase in all outputs. If the technology exhibits NDRTS, then scaling up (down) of all outputs and inputs leads to an increase (decrease) in the IDF. Likewise, if the technology exhibits NIRTS, then scaling up (down) of all outputs and inputs leads to a decrease (increase) in the IDF.

We can use (6) and (7) to quantify the magnitude of RTS. Using the cost-based RTS measure in (6), an increase in output quantities by a factor $\gamma > 1$ leads to an increase in cost by a factor $S^{\text{cost}}(\gamma|y, w) \times \gamma$. Conversely, a decrease in output quantities by a factor γ^{-1} leads to a decrease in cost by a factor $(S^{\text{cost}}(\gamma|y, w) \times \gamma)^{-1}$. Likewise, using the IDF-based measure of RTS in (7), an increase in output quantities by a factor $\gamma > 1$ requires an increase in input quantities by a factor $\gamma / S^{\text{IDF}}(\gamma|y, x)$. Conversely, a decrease in output quantities by a factor γ^{-1} requires a decrease in input quantities by a factor $S^{\text{IDF}}(\gamma|y, x)/\gamma$.[6]

Using the two-inputs one-output case, Figs. 1 and 2 illustrate how our proposed measure in (7), based on the IDF, can identify the nature and quantify the magnitude of RTS. Figure 1 illustrates the concept of the IDF. The IDF, $D_{\text{I}}(y, x)$, is given by the ratio β/α and represents the maximum scaling factor by which one needs to divide an arbitrary vector of inputs $x = (x_1, x_2) \in L(y)$ along the ray through x, so that the resulting vector x^* still belongs to $L(y)$. By definition of the IDF, $x/D_{\text{I}}(y, x)$ is contained in $L(y)$, but no point southwest of it is in $L(y)$.

In Fig. 2, we draw several isoquants corresponding to different output levels. For simplicity, we consider the nature of RTS along the ray $\mathcal{R}_0$ for which $x_1 = x_2$, but any other ray can be used. To produce $y = 1$, $y = 2$, and $y = 3$, we require $x = (1, 1)$, $x = (1.5, 1.5)$, and $x = (2, 2)$, respectively, indicating increasing RTS. Increasing output from $y = 3$ to $y = 3.75$, a 25 % increase requires increasing x by 25 %, indicating constant RTS. A similar reasoning applies up to $y = 4.75$. However, going from $y = 4.5$ to $y = 5$, a 11.11 % increase requires a 33.33 % increase in x, from $x = (3, 3)$ to $x = (4, 4)$, indicating decreasing RTS.

Now, for any arbitrary $x \in L(y)$, we can compute our scale economy measure in (7) along the ray $\mathcal{R}_0$. Note that along $\mathcal{R}_0$, β and α are easily computed and so is $D_{\text{I}}(y, x)$. Consider $x = (8, 8)$. Along $\mathcal{R}_0$, the IDF associated with $y = 1$ is $D_{\text{I}}(y = 1, x) = 8$. Likewise, $D_{\text{I}}(y = 2, x) = 16/3$. Thus, in this case, $S^{\text{IDF}}(\gamma = 2|y = 1, x) = (16/3)/(1/2 \times 8) = 4/3$, indicating IRTS as expected. Thus, doubling y from $y = 1$ to $y = 2$ requires only an increase of 50 % in all inputs. Similarly, $D_{\text{I}}(y = 3, x) = 4$. Thus, considering increasing y from $y = 2$ to $y = 3$ yields $S^{\text{IDF}}(\gamma = 3/2|y = 2, x) = (4)/(3/2 \times 4/3) = 2$, indicating IRTS, since this only requires an increase of 1/3 in inputs. For $y = 3.75$, $D_{\text{I}}(y = 3.75, x) = 16/5$. Thus, increasing y from $y = 3$ to $y = 3.75$ yields $S^{\text{IDF}}(\gamma = 1.25|y = 3, x) = (16/5)/(4/5 \times 4) = 1$, indicating CRTS. The same results follow for $S^{\text{IDF}}(\gamma = 1.2|y = 3.75, x) = 1$. For $y = 5$, $D_{\text{I}}(y = 5, x) = 2$. Thus, $S^{\text{IDF}}(\gamma = 10/9|y = 4.5, x) = (2)/(9/10 \times 8/3) = 5/6 < 1$, indicating DRTS. Then, increasing y from $y = 4.5$ to $y = 5$ requires an increase in inputs of $\gamma / S^{\text{IDF}}(\gamma = 10/9|y = 4.5, x) = (10/9)/(5/6) = 4/3$ which is greater than

[6] We show later [see eq. (9)] that the IDF-based measures of RTS can be interpreted in terms of cost as well. However, the IDF-based measure is more general, since it does not require imposition of the cost minimization assumption upon which (6) is defined.

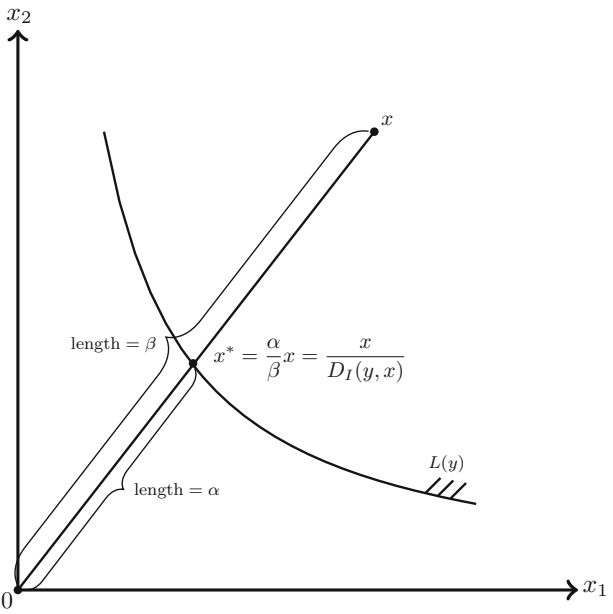

Fig. 1 Input distance function

the increase in output of 10/9. An analogous reasoning yields DRTS around $y = 5.5$ and $y = 7$.

Figure 2 also illustrates how our measure differs from the cost-based measure defined in (6). The cost-based measure assumes that isocost lines are tangent to the isoquant curves at each level of output y. Thus, it assumes that the firm is cost minimizing at the observed output level y and at all the hypothesized scaled output levels $\gamma\, y$. In addition, one needs input price and total cost information to determine the nature and magnitude of RTS using the cost-based measure. In contrast, using the IDF-based measure one can dispense with the assumption of cost minimization at each level of output and with input price data. In addition, the IDF-based measure is robust to the presence of technical inefficiency.[7]

To our knowledge, neither the measure of RTS based on the IDF that we propose nor the preceding analysis has appeared previously in the literature. In general, our measure of RTS based on the IDF can be used to compute RTS along any ray from the origin through any input vector. In particular, note that no cost-minimizing assumption is involved in the analysis. Moreover, the technology may exhibit varying degrees of RTS along different rays, and our measure based on the IDF will still be able to identify them. Another advantage of our approach is that we can measure RTS when outputs are scaled up or down by discrete factors and not at infinitesimal changes as traditional elasticity-based measures of RTS do.

[7] We later show that it is well defined even when $D_{\mathrm{I}}(y, x) > 1$.

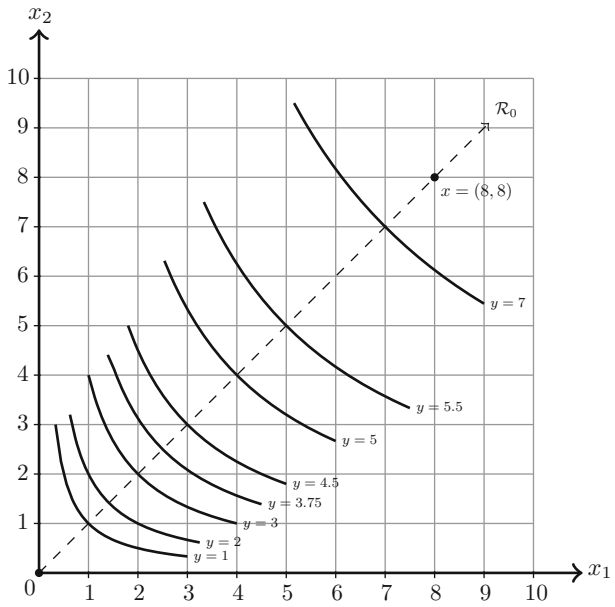

Fig. 2 Returns to scale

2.3 The related literature

Under the assumption of cost minimization, the cost function and the IDF are related through $wx = C(y, w)\, D_{\mathrm{I}}(y, x)$ [see Färe (1988), p 152]. Thus, the following equality holds:[8]

$$S^{\mathrm{cost}}(\gamma|y, w) = \frac{1}{S^{\mathrm{IDF}}(\gamma|y, x)} \tag{9}$$

W&W refer to $S^{\mathrm{cost}}(\gamma|y, w)$ in (6) as *ray-scale economies*. However, instead of using (6) directly, they base their analysis on how it changes when γ changes. In addition, they define an alternative measure of scale economies, viz., EPSE, which is in fact simply the ratio of two RSEs defined in (6). The EPSE in W&W is defined as

$$S^{\mathrm{W\&W}}(\theta|y, w) = \frac{C(\theta(1-\gamma)y, w)}{\theta C((1-\gamma)y, w)} \tag{10}$$

where $\theta = (1+\gamma)/(1-\gamma)$. Then,[9]

[8] If RTS is measured based on elasticity formula, the relationship in (9) can be obtained directly from the duality between cost and transformation functions [see Caves et al. (1981)]. We come to this in Sect. 5.2.

[9] They set $\gamma = 0.05$ and thus $\theta = 1.105$.

$$
\begin{aligned}
S^{\mathrm{W\&W}}(\theta|y,w) &= \frac{C(\theta(1-\gamma)y,w)}{\theta C((1-\gamma)y,w)} = \frac{C((1+\gamma)y,w)}{\frac{1+\gamma}{1-\gamma}C((1-\gamma)y,w)} \\
&= \frac{C((1+\gamma)y,w)}{(1+\gamma)C(y,w)} \times \frac{(1-\gamma)C(y,w)}{C((1-\gamma)y,w)} \\
&= S^{\mathrm{cost}}((1+\gamma)|y,w) \times \frac{1}{S^{\mathrm{cost}}((1-\gamma)|y,w)}
\end{aligned} \tag{11}
$$

Thus, the EPSE measure of W&W is the ratio of two different RSEs.

According to W&W, the EPSE measure gives an indication of RTS for a particular bank moving along the path from the origin through its observed output vector y, starting at $(1-\gamma)\times y$ and continuing to $(1+\gamma)\times y$. Perhaps, the more appropriate interpretation of (10) is that it measures RTS when outputs are scaled up by $\theta > 1$, starting at the observed output level scaled by $(1-\gamma)$. So, it measures RTS when outputs are scaled up by θ from $(1-\gamma)\times y$ to $(1+\gamma)\times y$ which is different from measuring RTS at or around y. That is, it does not measure RTS at or around the observed output level. The reason is that if from $(1-\gamma)\times y$ to $(1+\gamma)\times y$ the technology uniformly exhibits NIRTS, NDRTS, or CRTS, this measure is well defined. However, it is possible that when y is scaled up by $(1+\gamma)$, the technology might exhibit NDRTS (NIRTS or CRTS) but when y is scaled down by $(1-\gamma)$, it might exhibit NIRTS (NDRTS). In such a case, the EPSE measure is not well suited.

To illustrate this point further, suppose that we want to measure RTS around an hypothetical observed output level $y = 3$ with $\gamma = 1/3$ in our previous example. In this case, $\theta = 2$. Measuring RTS using the EPSE implies going from $y = 2$ to $y = \theta \times 2 = 4$. As Fig. 2 shows, in this case we will conclude that the technology exhibits NDRTS. However, the main interest is to see what happens around $y = 3$, which is the observed output level. From Fig. 2, it is clear that the technology exhibits NDRTS above $y = 3$ and CRTS below $y = 3$. So, the EPSE measure gives an indication of the nature and magnitude of RTS going from $y = 2$ to $y = 4$ but not around $y = 3$. That is, it gives an indication of RTS between two hypothetical points above and below the observed output level $y = 3$ but not at or around it.

Our RTS measures are based on global properties of the underlying technology. Thus, they can be readily extended to other available methods for estimating distance functions. For example, directional distance function along a particular direction vector can also be used. Our estimated measures are local in the sense that they measure RTS at the chosen value of γ. Choosing different values of γ allows us to make inferences about the robustness of our procedure. If for a given observation, $\hat{S}^{\mathrm{IDF}}(\gamma|x_0)$ changes wildly as γ changes, it may indicate that the nonparametric estimate of the IDF is unstable and unable to estimate RTS precisely. We check the stability of our results in the empirical section.

Under the assumption of cost minimization, the *ray-scale economies* measure of RTS based on the IDF is equivalent to the measure based on the cost function. This allows us to infer returns to scale and investigate how changes in the level of outputs change total costs for a given banking organization even if input prices are unavailable.

Further, our measure is still valid if the cost minimization assumption fails to hold in the data.

In the next section, we present our nonparametric estimation strategy for (7) and show how to make inferences about scale economies.

3 A model of banks' production process

Consistent with the widely used intermediation approach of Sealey and Lindley (1977), we assume that banks' balance sheets capture the essential structure of banks' core business: (i) liabilities, together with physical and financial (equity) capitals, and labor, are inputs into the bank production process and (ii) assets, other than physical, are outputs. Further, we define off-balance sheet activities as an additional output. Liabilities include core deposits and purchased funds, while assets include loans and trading securities. Off-balance sheet activities include all revenue sources other than lending and securities trading. Therefore, banks use labor, physical capital, financial capital, and liabilities to produce loans, invest in financial assets, and facilitate other financial services.

To keep our results comparable with those in the literature and related works, we define the following output variables: consumer loans (y_1), real estate loans (y_2), business loans (y_3), securities (y_4), and off-balance sheet activities (y_5). The input variables are purchased funds (x_1), core deposits (x_2), and number of full-time equivalent employees (x_3). In addition, we include physical (x_4) and financial (equity) capitals (x_5) as inputs for the IDF and quasi-fixed inputs for the cost function. We also include time as an extra variable to account for shifts in the technology (technical change) and possible variations in RTS estimates over time.

We use three different samples covering the period between 2000 and 2010. We obtain data for BHC from the Federal Reserve end-of-year FR Y-9C reports and for commercial banks from end-of-year Call reports. Except for employees' information, all variables are nominal stock variables. We use the U.S. GDP implicit price deflator to transform all nominal variables to 2010 prices. Table 1 reports summary statistics of relevant variables in the three samples. Sample I includes 8,265 observations for 1,418 top-tier BHC with assets above $500 million. Sample II includes 10,229 observations for 1,640 commercial banks with assets above $500 million. Sample III includes sample I plus 1,287 observations for 262 independent commercial banks not owned by a reporting BHC with assets above $500 million.

4 Returns to scale estimation

To compute our measure of RTS given by (7), we estimate the IDF $D_{\mathrm{I}}(y, x)$ and then evaluate $D_{\mathrm{I}}(\gamma y, x)$ for $\gamma = \{0.95, 0.97, ..., 1.05\}$. Following W&W, we assume away inefficiency and use nonparametric kernel methods to avoid making an arbitrary assumption about the functional form of the underlying technology. After imposing the linear homogeneity property (in inputs) on the IDF and including time into $D_{\mathrm{I}}(y, x) = 1$, the estimating equation becomes:

Table 1 Data summary statistics

	Mean	SD	Min	Percentiles					Max
				5th	25th	Median	75th	95th	
Sample I: BHC									
TA	16,200	115,000	500.00	539.00	697.00	1,060.0	2,510.0	39,400	2,290,000
C	782.00	5,490.0	5.0490	23.065	33.122	51.366	120.00	1,920.0	140,000.0
y_1	1,460.0	12,200	0.0196	2.1665	11.677	31.935	93.191	1,620.0	238,000.0
y_2	4,320.0	25,600	0.2238	217.00	352.00	548.00	1,220.0	11,200	553,000.0
y_3	2,150.0	13,700	0.0067	22.042	64.335	129.00	310.00	4,520.0	232,000.0
y_4	6,280.0	55,400	8.6079	82.488	163.00	283.00	664.00	8,800.0	1,250,000
y_5	333.00	2,630.0	0.0210	0.9787	2.8997	6.1790	19.120	545.00	56,700.00
x_1	8,740.0	75,700	7.0557	86.885	166.00	277.00	686.00	13,900	1,910,000
x_2	5,880.0	35,400	0.5570	328.00	452.00	683.00	1,510.0	15,900	874,000.0
x_3	2.8338	18.325	0.0330	0.1130	0.1840	0.2760	0.5880	8.0260	410.0000
x_4	169.00	911.00	0.3252	6.1599	13.812	23.305	48.774	505.00	18,000.00
x_5	1,400.0	9,640.0	0.3740	36.647	59.035	93.168	219.00	3,600.0	234,000.0
w_1	0.0370	0.0140	0.0050	0.0170	0.0280	0.0360	0.0460	0.0590	0.2280
w_2	0.0200	0.0100	0.0000	0.0060	0.0120	0.0180	0.0260	0.0380	0.2600
w_3	0.0664	0.0266	0.0072	0.0438	0.0532	0.0610	0.0732	0.1022	0.9795
Sample II: commercial banks									
TA	9,080.0	69,000	500.00	523.00	649.00	961.00	2,050.0	18,000	1,790,000
C	389.00	2,860.0	6.0000	19.942	28.919	43.208	90.839	791.00	70,200.00
y_1	691.00	5,860.0	0.0051	1.9656	10.843	29.114	80.906	981.00	158,000.0
y_2	2,810.0	17,800	0.1639	195.00	323.00	481.00	955.00	6,120.0	484,000.0
y_3	1,780.0	14,100	0.0111	28.998	71.212	131.00	283.00	3,240.0	395,000.0
y_4	3,890.0	36,900	18.355	94.204	178.00	296.00	658.00	6,380.0	1,110,000
y_5	176.000	1,420.0	0.0360	1.7547	4.8507	8.7714	22.695	318.00	37,600.00
x_1	3,400.0	33,800	10.695	65.802	130.00	214.00	465.00	4,320.0	1,000,000
x_2	5,420.0	40,700	11.177	312.00	425.00	626.00	1,300.0	10,900	979,000.0
x_3	1.7411	10.990	0.0092	0.0918	0.1742	0.2633	0.5051	4.3110	231.0000
x_4	98.599	559.00	0.0000	4.2311	11.252	19.210	38.820	269.00	14,300.00
x_5	867.00	6,210.0	0.9580	41.965	59.080	90.513	196.00	1,780.0	171,000.0
w_1	0.0370	0.0140	0.0050	0.0170	0.0280	0.0360	0.0460	0.0590	0.2280
w_2	0.0200	0.0100	0.0000	0.0060	0.0120	0.0180	0.0260	0.0380	0.2600
w_3	0.0595	0.0237	0.0058	0.0365	0.0464	0.0552	0.0666	0.0957	0.8333
Sample III: BHC and independent commercial banks									
TA	14,900	107,000	500.00	537.00	690.00	1,050.0	2,480.0	41,300	2,290,000
C	714.00	5,110.0	5.0490	22.124	32.245	49.990	116.000	1,920.0	140,000.0
y_1	1,330.0	11,400	0.0051	1.7830	10.887	30.458	90.745	1,710.0	238,000.0
y_2	4,060.0	24,000	0.2020	208.00	343.00	531.00	1,150.0	11,800	553,000.0
y_3	2,020.0	12,800	0.0067	22.378	64.822	129.00	316.00	5,160	232,000.0

Table 1 continued

	Mean	SD	Min	Percentiles					Max
				5th	25th	Median	75th	95th	
y_4	5,820.0	51,700	8.6079	84.586	166.00	289.00	683.00	10,100	1,250,000
y_5	300.000	2,450.0	0.0210	0.9899	3.0708	6.4256	19.370	554.000	56,700.00
x_1	7,870.0	70,500	7.0557	81.675	159.00	273.00	681.00	13,300	1,910,000
x_2	5,630.0	33,300	0.5570	320.00	446.00	671.00	1,490.0	17,900	874,000.0
x_3	2.5956	17.085	0.0092	0.1030	0.1790	0.2700	0.5705	7.5000	410.0000
x_4	155.00	851.00	0.0000	5.1652	13.143	22.454	47.131	476.00	18,000.00
x_5	1,330.0	9,010	0.3740	37.297	59.214	93.711	225.00	3,900	234,000.0
w_1	0.0370	0.0140	0.0050	0.0170	0.0280	0.0360	0.0460	0.0590	0.2280
w_2	0.0200	0.0100	0.0000	0.0060	0.0120	0.0180	0.0260	0.0380	0.2600
w_3	0.0671	0.0282	0.0067	0.0432	0.0530	0.0613	0.0739	0.1047	0.9795

Notes: This table reports summary statistics for the banking organizations used in estimation. Data cover the period 2001–2010 and correspond to annual values as of end-of-quarter of each year. Sample I includes 8,265 observations for 1,418 top-tier BHC with assets above $500 million. Sample II includes 10,125 observations for 1,640 commercial banks with assets above $500 million. Sample III includes sample I plus 1,287 observations for 262 independent commercial banks not owned by a BHC with assets above $500 million. Nominal values are in millions of 2010 dollars. Total assets (TA) correspond to balance sheet values. Total costs (TC) equal interest and noninterest expenses from income statement data. The output variables are consumer loans (y_1), real estate loans (y_2), business loans (y_3), securities (y_4), and off-balance sheet output (y_5). The input variables are purchased funds (x_1), core deposits (x_2), and labor (x_3). x_4 and x_5 correspond to the value of physical and equity capitals, respectively

$$-\ln x_1 = \ln D_I(y, \tilde{x}, t) + \epsilon \tag{12}$$

where $\tilde{x} = \{\ln(x_2/x_1), \ln(x_3/x_1), ..., \ln(x_5/x_1)\}$, $y = \{\ln y_1, \ln y_2, ..., \ln y_5\}$, $t = \{1, 2, ..., 11\}$ index time periods from 2001 to 2010, and the random error term, ϵ, captures the stochastic nature of the IDF.

Advances in nonparametric econometrics allow us to estimate (12) by smoothing over both the continuous (y and $\tilde{x}$) and the ordered (t) variables. To do this, we use a generalized kernel estimation along the lines of Li and Racine (2004) and Racine and Li (2004). Specifically, we use a local linear least- squares (LLLS) estimator which estimates a locally weighted least-squares regression around a point x_0 with weights determined by a kernel function and a bandwidth vector. Observations closer to x_0 receive a heavier weight. We use a second-order Gaussian kernel and least-squares cross validation (LSCV) to choose the bandwidths. We use the Wang and Ryzin (1981) bandwidth for the ordered variable. Hall et al. (2007) show that LSCV has desirable properties like the ability to smooth away irrelevant variables or to detect if some variables enter the equation linearly.

Our analysis so far has been based on the assumption that banks are fully efficient. This makes the nonparametric estimation simpler and is followed in the banking literature that uses nonparametric methods. If one assumes input-oriented inefficiency which is the norm in the parametric IDF literature, then the IDF can be expressed as

$$-\ln x_1 = \ln D_{\mathrm{I}}(y, \tilde{x}, t) - u + \epsilon \tag{13}$$

where $u \geq 0$ is input-oriented inefficiency. We can rewrite (13) as

$$-\ln x_1 = \ln \tilde{D}_{\mathrm{I}}(y, \tilde{x}, t) + \tilde{\epsilon} \tag{14}$$

where $\ln \tilde{D}_{\mathrm{I}}(y, \tilde{x}, t) = \ln D_{\mathrm{I}}(y, \tilde{x}, t) - \delta$, $\delta = E(u)$ and $\tilde{\epsilon} = \epsilon - [u - \delta]$. Although, by construction, $\tilde{\epsilon}$ has zero mean, it will estimate $\ln \tilde{D}_{\mathrm{I}}(y, \tilde{x}, t)$ instead of $\ln D_{\mathrm{I}}(y, \tilde{x}, t)$. Thus, from (14) we have $D_{\mathrm{I}}(y, \tilde{x}, t) \times \exp(-\delta) \neq D_{\mathrm{I}}(y, \tilde{x}, t)$. Since RTS in (8) is defined as the ratio of two IDF, and estimates of both of them are biased by the same multiplicative factor $\exp(-\delta)$, the ratio of the IDFs in (8) and therefore the RTS will be unaffected by the presence of inefficiency.[10] For these reasons, we do not introduce inefficiency into the model.

After estimating (12), we compute $\hat{D}_{\mathrm{I}}(q_0)$ which gives an estimate of $D_{\mathrm{I}}(y, \tilde{x}, t)$ at every point in $q_0 = (y, \tilde{x}, t)$—the vector of regressors. To compute $S^{\mathrm{IDF}}(\gamma|q_0)$, given in (7), we evaluate $\hat{D}_{\mathrm{I}}(q)$ for $\gamma = \{0.95, 0.97, ..., 1.05\}$ at every point in the vector of regressors $q = (\gamma\, y, \tilde{x}, t)$—note that γ is not a parameter to be estimated. For $\gamma = 1$, we obtain the estimated values of $D_{\mathrm{I}}(y, \tilde{x}, t)$ which enter in the denominator of (12). Using different values of $\gamma \in [0.95, 1.05]$ allows us to estimate RTS between the 95 % and the 105 % of the observed output vector for each banking organization. The special case considered by W&W corresponds to computing the ratio between $S^{\mathrm{IDF}}(1.05|q_0)$ and $S^{\mathrm{IDF}}(0.95|q_0)$.

To determine if a given observation is consistent with increasing, decreasing, or constant RTS, we need to compute the confidence interval (CI) for each point estimate of our RTS measures. A given value of $\hat{S}^{\mathrm{IDF}}(\gamma|q_0)$ indicates increasing RTS (IRTS) if it exhibits NDRTS and not CRTS. It indicates decreasing RTS (DRTS) if it exhibits NIRTS and not CRTS. It indicates constant RTS (CRTS) if it exhibits neither NIRTS nor NDRTS.

To construct bias-corrected CI, we use the wild bootstrap procedure. We bootstrap $S^{\mathrm{IDF}}(\gamma|q_0)$ for each cross section 399 times. Then we compute $\hat{D}_I(q)$ for $\gamma = \{0.95, 0.97, ..., 1.05\}$ 399 times, taking their associated residuals, drawing new residuals from a two-value distribution, as detailed in Härdle and Mammen (1993), and computing $S^{\mathrm{IDF}}(\gamma|q_0)$ again. After this, we estimate the standard errors, the bias of $\hat{S}^{\mathrm{IDF}}(\gamma|q_0)$, and its associated 95 % asymptotic CI.

For $\gamma > 1$, a given observation exhibits IRTS if the lower bound of its associated CI lies above 1. It exhibits CRTS if the lower bound of its associated CI lies below 1 and its upper bound above 1. It exhibits DRTS if the upper bound of its associated CI lies below 1. For $\gamma < 1$, the reverse is true. That is, a given observation exhibits IRTS if the upper bound of its associated CI lies below 1. It exhibits CRTS if the lower bound of its associated CI lies below 1 and its upper bound above 1. It exhibits DRTS if the lower bound of its associated CI lies above 1.

[10] This is also true when RTS is defined in terms of the elasticity-based formula in Sect. 5.2, viz., $RTS^{-1} = -\sum_m \partial \ln x_1 / \partial \ln y_m$ which is not affected by the presence of the inefficiency term u.

5 Empirical results

5.1 Ray-scale economies

We estimate the model in (12) as described in Sect. 4 for samples I, II, and III. We obtain similar results across the three samples. Table 2 presents the bandwidths estimates, their associated scale factors, residual standard errors, values of cross validation objective functions (CVOF), and R^2 values for the three samples. All variables are treated as continuous variables except for t which is treated as an ordered variable.[11] The R^2 values indicate a good fit of the model in (12) for the three samples. The estimated bandwidths differ across the three samples, but their variation seems small. In addition, the estimated bandwidth is small indicating that all variables are relevant in the empirical model.[12]

Using the estimated models for the three different samples, we compute our RSE measure based on the IDF using (7) and their corresponding CIs as described in the previous section. Table 3 reports[13] the median values of RSE estimates by size quartiles based on total assets for $\gamma = \{0.95, 0.97, ..., 1.05\}$. For $\gamma < 1$ a median value of RSE less (greater) than 1 indicates IRTS (DRTS). For $\gamma > 1$, a median value of RSE greater (less) than 1 indicates IRTS (DRTS). If the median value equals 1, it indicates CRTS. Thus, according to the median values of RSE for each quartile and each value of γ in Table 3, the median banking organization in each sample exhibits IRTS.

To determine if a given bank-year observation exhibits IRTS, CRTS or DRTS, we construct bias-corrected CIs for RSE as described in the previous section. The last three columns for each sample show the percentage of observations in each size quartile (based on total assets) that exhibit IRTS, CRTS, or DRTS. For instance, for Sample I and $\gamma = 0.95$, 47 % of observations in the first quartile show evidence of IRTS, 40.84 % CRTS, and 12.16 % DRTS. For the fourth quartile, the corresponding values are 43.51 %, 26.89 %, and 29.60 %. Overall, for $\gamma = 0.95$, 41.40 % of the observations show evidence of IRTS, 37.49 % CRTS, and 29.60 % DRTS. Table 3 shows that, compared to BHCs (Sample I), we find stronger evidence of IRTS for commercial banks (Sample II).

[11] We use the parallel implementation of the np-package in R, Hayfield and Racine 2008, for our estimations. The R^2 values are computed as the square of the correlation coefficient between the left-hand-side variable in (12) and its nonparametric estimate.

[12] As mentioned in Sect. 2, the IDF has to satisfy some theoretical properties. In particular, the IDF is linearly homogeneous in x, nondecreasing in x, and nonincreasing in y. The first property is imposed before estimation so it is satisfied at each data point. The other two properties require that $\partial D(y,x)/\partial x \geq 0$ and $\partial D(y,x)/\partial y \leq 0$ at each observation. However, in practice these properties are rarely satisfied at every data point, and it is nontrivial to impose these constraints when the the number of observations is large [see Du et al. (2013) and Hall and Huang (2001) for such a procedure]. We compute the percentage of violations of these theoretical properties in our IDF estimates. For the outputs (y_1 - y_5) the average percentage of violations across the three samples is 3.62, 0.60, 1.84, 0.27, and 3.41, respectively. For the inputs (x_2 - x_5) the average percentage of violations across the three samples are 0.62, 2.00, 3.34, and 3.91, respectively. Given that the percentage of violations is relatively small, we decided not to impose these constraints in the present paper.

[13] We dropped 9 observations from sample I, 14 observations from sample II, and 20 observations from sample III for which the estimated RSE was highly implausible.

Table 2 Nonparametric regression results from IDF

Variable	Sample I		Sample II		Sample III	
	Bandwidth	Scl. factor	Bandwidth	Scl. factor	Bandwidth	Scl. factor
$\ln x_2$	1.1292	4.068889	0.9276	3.045643	1.3186	4.596545
$\ln x_3$	0.3261	1.059223	0.3607	1.049982	0.3378	1.059479
$\ln x_4$	0.4012	1.059223	1.2462	3.087151	0.4194	1.059535
$\ln x_5$	0.2874	1.059223	0.2980	1.059223	0.2948	1.059273
$\ln y_1$	1.8149	2.359051	0.7869	1.071292	0.8230	1.060344
$\ln y_2$	0.8967	1.950924	1.4508	3.672538	1.7069	3.850792
$\ln y_3$	0.6171	1.059281	1.7517	3.466948	0.6142	1.059332
$\ln y_4$	1.5634	3.007355	1.1242	2.351977	0.7378	1.426104
$\ln y_5$	0.7399	1.059104	0.5949	1.059223	0.7143	1.059382
t	0.5000	1	0.5000	1	0.5000	1
Res. S.E.	0.0041		0.0017		0.0076	
CVOF	0.1700		0.0122		3.7996	
R^2	0.9983		0.9991		0.9969	
Obs.	8,265		10,125		9,550	

Notes: This table shows the local linear least-squares nonparametric regression bandwidths estimates and the associated scale (Scl.) factors using least-squares cross validation. All variables are treated as continuous except for t which is treated as an ordered variable. We use second-order Gaussian kernels for the continuous variables and an ordered categorical kernel for the ordered variable. Estimations are done using the parallel implementation of the np-package in R, Hayfield and Racine (2008). Res. S.E. stands for residuals standard error and CVOF for cross validation objective function. The R^2 values are computed as the squared of the correlation coefficient between the left-hand-side variable in (12) and its nonparametric estimate

The most salient result in Table 3 is that not all banking organizations exhibit IRTS. This result is different from W&W and holds across the three different samples. Another important result is that the number of observations indicating DRTS increases with total assets. For samples I and III, the number of observations indicating IRTS decreases as total assets increase up to the third quartile. However, the number of observations indicating IRTS tends to be higher for the largest banking organizations, those in the 4th quartile. In contrast, for sample II, which includes commercial banks only, the number of observations indicating IRTS decreases as total assets increase.

The results in Table 3 offer a more detailed description of the nature of RTS around the observed output and input levels. For each banking organization, starting at 95 % and continuing until 105 % of its observed output levels, the table shows that RSE measures change depending on the scaling factor γ by which outputs are scaled up or down. However, for different values of γ, the number of observations indicating IRTS, CRTS, or DRTS varies only slightly across all size quartiles. This indicates that around ± 5 % of the observed output levels, our measure of RTS is robust and stable. In Sample I, for $\gamma = 0.95$, 41.40 % of the observations indicate IRTS, while for $\gamma = 1.05$ this number is only 39.51 %. Thus, some observations (156) show IRTS when outputs are scaled down to 95 % of their observed level but CRTS or DRTS when outputs are scaled up to 105 % of their observed level. Similar results hold true for samples II and III. In these cases, the EPSE measure in (10) will be inconclusive.

Table 3 Ray-scale economies (RSE) summary statistics

Quartile	Sample I: BHC					Sample II: commercial banks					Sample III: BHC and ind. comm. banks				
	Obs.	Median	IRTS	CRTS	DRTS	Obs.	Median	IRTS	CRTS	DRTS	Obs.	Median	IRTS	CRTS	DRTS
$\gamma = 0.95$															
1	2,064	0.9988	47.00	40.84	12.16	2,528	0.9972	72.11	23.66	4.23	2,383	0.9991	40.91	36.26	22.83
2	2,064	0.9993	40.41	41.52	18.07	2,528	0.9979	65.82	28.96	5.22	2,382	0.9995	38.08	35.05	26.87
3	2,064	0.9995	34.69	40.70	24.61	2,528	0.9985	52.69	37.10	10.21	2,383	0.9997	34.49	37.18	28.33
4	2,064	0.9995	43.51	26.89	29.60	2,527	0.9987	50.49	27.74	21.76	2,382	0.9996	44.67	25.06	30.27
	8,256		41.40	37.49	21.11	10, 111		60.28	29.36	10.36	9,530		39.54	33.39	27.07
$\gamma = 0.97$															
1	2,064	0.9993	46.27	41.04	12.69	2,528	0.9984	71.36	24.01	4.63	2,383	0.9995	40.83	36.09	23.08
2	2,064	0.9996	39.92	41.52	18.56	2,528	0.9988	65.15	29.71	5.14	2,382	0.9997	36.57	35.85	27.58
3	2,064	0.9997	34.59	40.65	24.76	2,528	0.9991	52.69	37.18	10.13	2,383	0.9998	34.20	37.39	28.41
4	2,064	0.9997	43.07	27.33	29.60	2,527	0.9992	50.26	28.29	21.45	2,382	0.9997	44.25	25.15	30.60
	8,256		40.96	37.63	21.40	10,111		59.87	29.80	10.34	9,530		38.96	33.62	27.42
$\gamma = 0.99$															
1	2,064	0.9998	45.83	41.47	12.69	2,528	0.9995	71.12	24.25	4.63	2,383	0.9999	39.99	36.30	23.71
2	2,064	0.9999	38.71	42.10	19.19	2,528	0.9996	64.99	29.59	5.42	2,382	0.9999	36.06	35.77	28.17
3	2,064	0.9999	34.06	40.94	25.00	2,528	0.9997	52.10	37.58	10.32	2,383	0.9999	34.28	37.43	28.28
4	2,064	0.9999	43.51	26.99	29.51	2,527	0.9998	50.22	28.22	21.57	2,382	0.9999	43.79	25.65	30.56
	8,256		40.53	37.88	21.60	10,111		59.61	29.91	10.48	9,530		38.53	33.79	27.68

Table 3 continued

Quartile	Sample I: BHC					Sample II: commercial banks					Sample III: BHC and ind. comm. banks				
	Obs.	Median	IRTS	CRTS	DRTS	Obs.	Median	IRTS	CRTS	DRTS	Obs.	Median	IRTS	CRTS	DRTS
$\gamma = 1.01$															
1	2,064	1.0002	45.30	41.62	13.08	2,528	1.0005	70.89	24.37	4.75	2,383	1.0002	39.19	36.72	24.09
2	2,064	1.0001	37.94	42.34	19.72	2,528	1.0004	64.16	30.50	5.34	2,382	1.0001	35.47	35.94	28.59
3	2,064	1.0001	34.06	40.26	25.68	2,528	1.0003	52.06	37.42	10.52	2,383	1.0001	33.95	37.26	28.79
4	2,064	1.0001	43.12	27.42	29.46	2,527	1.0002	50.34	27.86	21.80	2,382	1.0001	43.95	25.06	30.98
	8,256		40.10	37.91	21.98	10,111		59.36	30.04	10.60	9,530		38.14	33.75	28.11
$\gamma = 1.03$															
1	2,064	1.0006	44.53	41.57	13.91	2,528	1.0015	70.09	24.88	5.02	2,383	1.0004	38.65	36.30	25.05
2	2,064	1.0004	37.55	42.10	20.35	2,528	1.0011	64.00	30.66	5.34	2,382	1.0002	34.93	36.06	29.01
3	2,064	1.0003	34.06	40.16	25.78	2,528	1.0008	52.57	36.63	10.80	2,383	1.0002	33.53	37.31	29.16
4	2,064	1.0003	43.12	27.23	29.65	2,527	1.0007	50.14	28.45	21.41	2,382	1.0003	44.29	24.60	31.11
	8,256		39.81	37.77	22.42	10,111		59.20	30.16	10.64	9,530		37.85	33.57	28.58
$\gamma = 1.05$															
1	2,064	1.001	43.99	41.76	14.24	2,528	1.0024	69.66	25.20	5.14	2,383	1.0007	37.94	36.59	25.47
2	2,064	1.0006	37.11	41.62	21.27	2,528	1.0018	63.25	31.45	5.30	2,382	1.0003	34.01	36.15	29.85
3	2,064	1.0004	34.16	39.87	25.97	2,528	1.0013	52.37	36.55	11.08	2,383	1.0003	33.40	37.35	29.25
4	2,064	1.0005	42.78	27.71	29.51	2,527	1.0012	50.02	28.33	21.65	2,382	1.0004	44.12	24.56	31.32
	8,256		39.51	37.74	22.75	10,111		58.83	30.38	10.79	9,530		37.37	33.66	28.97

Notes: This table shows the median values of RTS estimates for different values of γ by size quartiles based on total assets and the percentage of observation exhibiting increasing, constant, and decreasing returns to scale (IRTS, CRTS, and DRTS, respectively). For $\gamma < 1$ ($\gamma > 1$), a median value of RTS less (greater) than one indicates IRTS. For $\gamma > 1$ ($\gamma < 1$), a median value of RTS greater (less) than one indicates DRTS. If the median value equals one, it indicates CRTS. The last three columns for each sample show the percentage of observations in each quartile for which the bias-corrected wild bootstrap CIs indicate IRTS, CRTS or DRTS

Table 4 Summary statistics of input response for a proportional change ($\gamma > 0$) in all outputs

γ	Mean	SD	Min	Percentiles					Max
				5th	25th	Median	75th	95th	
Panel A. Sample I: BHC									
0.95	0.9511	0.0066	0.6185	0.9480	0.9498	0.9507	0.9517	0.9550	1.1808
0.97	0.9707	0.0029	0.9178	0.9688	0.9699	0.9704	0.9710	0.9730	1.0572
0.99	0.9902	0.0012	0.9788	0.9896	0.9899	0.9901	0.9903	0.9910	1.0549
1.01	1.0098	0.0011	0.9652	1.0090	1.0097	1.0099	1.0101	1.0104	1.0234
1.03	1.0293	0.0036	0.8668	1.0270	1.0290	1.0296	1.0302	1.0313	1.0521
1.05	1.0488	0.0073	0.6915	1.0449	1.0483	1.0494	1.0503	1.0521	1.0866
Panel B. Sample II: commercial banks									
0.95	0.9521	0.0042	0.8575	0.9486	0.9507	0.9518	0.9530	0.9556	1.0879
0.97	0.9712	0.0025	0.9124	0.9691	0.9704	0.9711	0.9718	0.9733	1.0305
0.99	0.9904	0.0008	0.9701	0.9897	0.9901	0.9904	0.9906	0.9911	1.0100
1.01	1.0096	0.0008	0.9902	1.0089	1.0094	1.0096	1.0099	1.0103	1.0307
1.03	1.0291	0.0156	0.6327	1.0266	1.0282	1.0289	1.0296	1.0309	1.8444
1.05	1.0482	0.0164	0.6446	1.0443	1.0470	1.0482	1.0493	1.0515	1.8795
Panel C. Sample III: BHC and independent commercial banks									
0.95	0.9511	0.0050	0.8850	0.9472	0.9494	0.9505	0.9518	0.9558	1.0518
0.97	0.9706	0.0030	0.9454	0.9683	0.9696	0.9703	0.9711	0.9735	1.0305
0.99	0.9902	0.0010	0.9816	0.9894	0.9899	0.9901	0.9904	0.9912	1.0100
1.01	1.0098	0.0010	0.9903	1.0088	1.0097	1.0099	1.0101	1.0106	1.0186
1.03	1.0294	0.0031	0.9713	1.0265	1.0290	1.0297	1.0304	1.0318	1.0562
1.05	1.0490	0.0054	0.9503	1.0441	1.0483	1.0496	1.0507	1.0531	1.0947

Notes: This table shows the proportion by which input quantities increase or decrease when all outputs are multiplied by $\gamma > 0$ (in column one). Panels A, B, and C, show results for Samples I, II, and III, respectively. For example, for Sample I, multiplying all outputs by $\gamma = 1.05$ requires an average increase in all input quantities by 1.0488 which is less than 1.05, indicating increasing returns to scale. Likewise, multiplying all output quantities by $\gamma = 0.95$ leads to an decrease in all input quantities, on average, by 0.9511 which is greater than 0.95. This, again, indicates the presence of increasing returns to scale

Table 4 presents summary statistics of the percentage by which input quantities will change when all outputs are multiplied by a factor $\gamma > 0$. Panel A, B, and C, show results for samples I, II, and III, respectively. For sample I, multiplying all outputs by $\gamma = 1.05$ requires an average increase in all input quantities by a factor of 1.0488 which is less than 1.05, indicating IRTS. Alternatively, multiplying all output quantities by $\gamma = 0.95$ requires an average decrease in all input quantities by a factor of 0.9511 which is greater than 0.95, indicating IRTS as well. Under cost minimization, the factors presented in Table 4 represent the factors by which total costs will increase or decrease when all outputs are scaled up or down by γ.

Table 4 shows that, despite the statistical evidence indicating the presence of economies of scale for some of the biggest BHC and commercial banks in the U.S. as presented in Table 3, RTS seem to be economically small for the range of output changes we consider. Regardless of the value of γ, the median RSE estimate for all

three samples is around 1, indicating nearly CRTS for all observations. Less than 1 % of the observations show RSE measures consistent with economically significant RTS. For instance, only around 34 observations indicate economies of scale greater than 1.03, implying that a 10 % increase in all outputs will require an increase in all inputs by approximately 9.7 % ($1/1.03 \times 10\%$). For all other observations, an increase of 10 % in all outputs will require an increase in all inputs by approximately the same percentage. Thus, despite the existence of scale economies for some large BHC and commercial banks, these economies of scale seem to be economically small.

For completeness, we estimate the cost model of Wheelock and Wilson (2012) using Sample I for BHC. They consider the same output and input variables we use for the IDF. However, since no accurate input prices exist for physical and financial capitals, they consider them as quasi-fixed inputs. Table 5 reports the nonparametric regression bandwidth estimates for this model. Table 6 presents the median RSE estimates using (6) and the number of observations exhibiting IRTS, CRTS, or DRTS. For comparability with our IDF results, we do not use any dimension reduction technique to estimate W&W's model. Thus, the results are readily comparable with those presented in Table 3 for our IDF-based RSE estimates. Using the cost function, we find more evidence in favor of IRTS. However, we still find observations indicating constant or decreasing RTS. Overall, about 73 % of the observations show evidence of IRTS, 20 % indicate CRTS, and the remaining 7 % indicate DRTS.

In terms of the economic significance of the RTS estimates from the cost function, our results indicate that despite the existence of IRTS, economies of scale seem to be small. Table 7 presents summary statistics of the proportion by which total costs would change if outputs were scaled up or down by γ. For instance, multiplying all outputs by $\gamma = 1.05$ leads to an average increase in total cost by approximately 4.32 % which is less than the 5 % increase in all outputs, indicating IRTS. Likewise, multiplying all output quantities by $\gamma = 0.95$ leads to an average decrease in total cost by about 5.66 % which is more than the 5 % decrease in all outputs. This, again, indicates IRTS. Overall, the results from both the IDF and the cost function show that economies of scale in the banking industry seem to be small and that there are some banking organizations that operate under constant or decreasing returns to scale.

The differences between our RTS estimates from the IDF and the cost function are not surprising, since these two approaches use different data. Moreover, RTS estimates from the cost function should be interpreted as short run RTS, since the cost model includes physical and financial capitals as quasi-fixed inputs. In contrast, RTS estimates from the IDF assume that physical and financial capitals are variable inputs—a more appropriate assumption for the long run. In any case, our evidence does not support the view that all U.S. banking organizations exhibit IRTS and, given that the scale economies seem to be small, breaking up some of the largest banking organization may not lead to substantial welfare losses.

5.2 Elasticity-based returns to scale

A traditional measure of RTS is based on cost elasticities with respect to each output and is defined as

Table 5 Cost function nonparametric regression estimates

Variable	Bandwidth	Scale factor
$\ln w_1$	3.8656	16.5420
$\ln w_2$	1.2083	3.8636
$\ln y_1$	0.8149	1.0592
$\ln y_2$	1.4868	3.2350
$\ln y_3$	0.6164	1.0580
$\ln y_4$	0.5507	1.0592
$\ln y_5$	0.7400	1.0592
$\ln x_4$	0.4950	1.0592
$\ln x_5$	0.5141	1.0592
t	0.5000	1
Residual S.E.	0.0031	
R^2	0.9982	
CVOF	0.0141	
Observations	8,265	

Notes: This table shows the local linear least-squares nonparametric regression bandwidths estimates using least-squares cross validation for Wheelock and Wilson (2012)'s cost model using data for BHC (Sample I). The left-hand-side variable equals $\ln C/w_3$. All variables are treated as continuous variables except for t which is treated as an ordered variable. We use second-order Gaussian kernels for the continuous variables and an ordered categorical kernel for the ordered variable. Estimations are done using parallel implementation of the np-package in R, Hayfield and Racine (2008). S.E. is standard error, CVOF is cross validation objective function. The R^2 value is computed as the squared of the correlation coefficient between the left-hand-side variable in (12) and its nonparametric estimate

$$\eta(w, y) = \left(\sum_l \frac{\partial \ln C(w, y)}{\partial \ln y_l} \right)^{-1} \tag{15}$$

Alternatively, using an IDF, this measure is equivalent to[14]

$$\eta(x, y) = -\left(\sum_l \frac{\partial \ln D(x, y)}{\partial \ln y_l} \right)^{-1} \tag{16}$$

Values of $\eta(\cdot) \gtreqless 1$ indicate increasing, constant, or decreasing RTS, respectively. To compute (15) ((16)), one needs to estimate the derivatives of the cost function (IDF). W&W argue that estimates of the derivatives of nonparametric cost function

[14] This elasticity-based formula is a special case of the formula based on the transformation function $F(x, y) = 1$ for which RTS is defined as $RTS(x, y) = -\sum_j \frac{\partial \ln F(\cdot)}{\partial \ln x_j} \div \sum_l \frac{\partial \ln F(\cdot)}{\partial \ln y_l}$ (Caves et al. 1981). Note that for IDF $\sum_j \frac{\partial \ln F(\cdot)}{\partial \ln x_j} = 1$, which follows from the linear homogeneity (in x) property of the IDF.

Table 6 Cost function RSE estimates for sample I

Quartile	Obs.	Median	IRTS	CRTS	DRTS	Median	IRTS	CRTS	DRTS
		$\gamma = 0.95$				$\gamma = 1.01$			
1	2066	1.0071	72.17	23.91	3.92	0.9933	73.72	22.41	3.87
2	2065	1.0078	79.66	17.53	2.81	0.9925	80.48	16.66	2.86
3	2066	1.0073	76.43	18.44	5.13	0.9933	76.09	18.39	5.52
4	2065	1.0055	64.02	21.79	14.19	0.9949	63.78	21.36	14.87
	8262		73.07	20.42	6.51		73.52	19.70	6.78
		$\gamma = 0.97$				$\gamma = 1.03$			
1	2066	1.0042	72.60	23.43	3.97	0.9959	73.33	22.75	3.92
2	2065	1.0047	79.61	17.48	2.91	0.9955	79.61	17.48	2.91
3	2066	1.0043	76.23	18.54	5.23	0.9959	75.99	18.30	5.71
4	2065	1.0032	64.02	21.55	14.43	0.9969	63.73	21.26	15.01
	8262		73.12	20.25	6.63		73.17	19.95	6.89
		$\gamma = 0.99$				$\gamma = 1.05$			
1	2066	1.0014	72.80	23.28	3.92	0.9933	73.48	22.41	4.11
2	2065	1.0016	79.66	17.48	2.86	0.9925	79.56	17.34	3.10
3	2066	1.0014	76.14	18.49	5.37	0.9933	76.19	18.05	5.76
4	2065	1.0011	63.92	21.55	14.53	0.9949	63.58	21.31	15.11
	8262		73.13	20.20	6.67		73.20	19.78	7.02

Notes: This table shows the median values of RTS estimates based on the cost function for different values of γ by size quartiles based on total assets. IRTS, CRTS, and DRTS stand for increasing, constant, and decreasing returns to scale. For $\gamma < 1$ ($\gamma > 1$), a median value of RTS greater (less) than one indicates IRTS. For $\gamma > 1$ ($\gamma < 1$), a median value of RTS less (greater) than one indicates DRTS. If the median value equals one, it indicates CRTS. The last three columns for each sample show the number of observations in each quartile for which the bias- corrected wild bootstrap CIs indicate IRTS, CRTS, or DRTS. We use 99 bootstrap replicates to construct the CIs

Table 7 Summary statistics of change in cost for a proportional change ($\gamma > 0$) in all outputs

γ	Mean	SD	Min	Percentiles					Max
				5th	25th	Median	75th	95th	
0.95	0.9434	0.0076	0.8746	0.9345	0.9407	0.9433	0.9461	0.9525	1.0823
0.97	0.9660	0.0046	0.9247	0.9606	0.9644	0.9659	0.9677	0.9716	1.0508
0.99	0.9887	0.0016	0.9744	0.9868	0.9881	0.9886	0.9892	0.9906	1.0173
1.01	1.0087	0.0016	0.9948	1.0068	1.0081	1.0086	1.0092	1.0106	1.0384
1.03	1.0259	0.0048	0.9781	1.0203	1.0242	1.0258	1.0277	1.0318	1.1212
1.05	1.0432	0.0082	0.8975	1.0338	1.0402	1.0430	1.0461	1.0531	1.2123

Notes: This table shows the proportion by which total costs would increase or decrease when all output quantities are changed by $\gamma > 0$ proportion (in column one). For example, for Sample I, multiplying all outputs by $\gamma = 1.05$ leads to an average increase in total costs by 1.0432 which is less than 1.05, indicating increasing returns to scale. Likewise, multiplying all output quantities by $\gamma = 0.95$ leads to an average decrease in total costs by 0.9434 which is less than 0.95. Again, this indicates the presence of increasing returns to scale

may be noisier than the estimate of the cost function itself. For this reason, they estimate RTS using (10) instead of (15). However, we think that the above concern is an empirical issue. We estimate (15) and (16) and compare the results with our previous RSE estimates.

We refer to (15) and (16) as *elasticity-based* RTS (EB-RTS). To obtain them, we need the estimated derivatives of the cost and input distance functions with respect to outputs. Fortunately, these derivatives are obtained as a by-product of the local linear least-squares nonparametric regression approach. Thus, the only additional task is to estimate the CIs.

We compute (15) for sample I and (16) for samples I, II, and III. Tables 8 and 9 report the results.[15] Panel A (B) of Table 8 shows the results for the EB-RTS estimates using the IDF (cost function). Again, we report separate results for each size quartile based on total assets. Overall, the results from the IDF for the three different samples closely mirror those reported in Table 3 in terms of percentage of observations consistent with IRTS, CRTS, and DRTS. For example, the results presented for Sample I in Table 3 show that, on average, 40.39, 37.74, and 21.88 % of the observations show evidence of IRTS, CRTS, and DRTS, respectively. The corresponding figures using the elasticity-based RTS estimates are 39.36, 40.80, and 19.84 % (Table 8, Panel A). The similarities also hold within each size quartile. The results are also similar for samples II and III. Thus, using the EB-RTS estimates, our previous results regarding the nature of economies of scale exhibited by the biggest banking organizations are essentially unchanged.

In terms of the EB-RTS estimates using the cost function, the results are also comparable with our previous results based on RSE and presented in Table 6. It can be seen from this table that 73, 20, and 7 % of the observations show evidence consistent with IRTS, CRTS, and DRTS, respectively. Table 8 shows that we obtain similar results using the EB-RTS estimates—75, 20, and 5 %, respectively. Again, our previous results remain qualitatively unchanged.

Table 8 shows that the median EB-RTS estimate using the IDF is close to 1 for sample I and III and slightly above 1 for sample II. These results are consistent with our previous finding that economies of scale seem to be economically small. Table 9 presents additional summary statistics for EB-RTS estimates. This table shows that, in general, the EB-RTS estimates are higher than their counterparts derived using (7).

An important difference is that EB-RTS estimates obtained using the cost function seem economically significant and higher than the corresponding values estimated using (6). Further, they seem to be higher than those estimated using the IDF. The differences may stem from the facts that (i) the cost minimization behavior is not imposed on the estimated IDF, and (ii) the IDF treats physical and financial capitals as variable inputs, while the cost function treats them as quasi-fixed inputs. Moreover, the cost function uses input price data, whereas IDF uses input quantities. To make the cost function and IDF results comparable, we adjust the cost function EB-RTS estimates by multiplying them by one minus the sum of the elasticities of the cost function with

[15] The last column of Table 9 presents the number of observations used in the computations.

Table 8 Scale economies from elasticity-based RTS estimates

Panel A: IDF estimates										
Quartile	Sample I: BHC					Sample II: commercial banks				
	Obs.	Median	IRTS	CRTS	DRTS	*Obs.*	Median	IRTS	CRTS	DRTS
1	2,041	1.0235	52.13	38.22	9.65	2,521	1.0506	77.67	20.67	1.67
2	2,056	1.0131	39.88	45.43	14.69	2,512	1.0372	73.61	24.32	2.07
3	2,060	1.0059	24.47	49.27	26.26	2,518	1.0257	60.76	35.15	4.09
4	2,024	1.0089	41.11	30.09	28.80	2,465	1.0234	52.94	29.21	17.85
Total	8,181	1.0119	39.36	40.80	19.84	10,016	1.0340	66.31	27.33	6.36
Quartile	Sample III: CB and BHC									
	Obs.	Median	IRTS	CRTS	DRTS					
1	2,373	1.0154	42.60	39.15	18.25					
2	2,377	1.0062	28.44	44.09	27.47					
3	2,378	1.0009	23.17	40.45	36.38					
4	2,326	1.0044	40.46	26.74	32.80					
Total	9,454	1.0060	33.63	37.67	28.71					
Panel B: cost function estimates										
Quartile	Sample I: BHC					Sample I—adjusted: BHC				
	Obs.	Median	IRTS	CRTS	DRTS	*Obs.*	Median	IRTS	CRTS	DRTS
1	2,049	1.1718	79.01	19.03	1.95	2, 045	1.0546	35.94	56.77	7.29
2	2,062	1.1754	80.26	17.94	1.79	2, 057	1.0487	30.77	64.46	4.76
3	2,063	1.1581	76.49	19.39	4.12	2, 060	1.0319	25.87	66.21	7.91
4	2,007	1.1152	64.42	22.47	13.10	2, 019	1.0154	31.70	49.13	19.17
Total	8,181	1.1614	75.11	19.69	5.19	8, 181	1.0401	31.06	59.20	9.74

Notes: This table shows the median values of elasticity-based return to scale (RTS) estimates by size quartiles based on total assets and the percentage of observation exhibiting increasing, constant, and decreasing returns to scale (IRTS, CRTS, and DRTS, respectively). Panels A and B report information for elasticity-based RTS estimates derived from nonparametric IDF and the nonparametric cost function, respectively. The last three columns for each sample show the percentage of observations in each quartile for which the bias- corrected wild bootstrap CIs indicate IRTS, CRTS, or DRTS

respect to both physical and financial capitals (see Caves et al. (1981)). The results after this adjustment are presented in Panel B of Table 9. This adjustment lowers the median estimate of RTS from 1.17 to 1.05, suggesting that the unadjusted EB-RTS are biased upward. Further, the adjusted EB-RTS estimates are closer to those estimated from the IDF.[16] Overall, we find that the EB-RTS estimates favor our earlier results

[16] By the LeChatelier Principle, assuming that physical and financial capitals are quasi-fixed inputs, total cost will respond slowly to changes in other variables compared to the case when all inputs are variable. Thus, measured elasticities are lower than what they would be if all inputs were variable. Consequently, measured RTS will tend to be higher, since they are computed as the inverse of a sum of elasticities (as shown in equations (15) and (16)).

Table 9 Summary statistics of elasticity-based RTS

Quartiles	Mean	SD	Min	Percentiles					Max	*N*
				5th	25th	Median	75th	95th		
Panel A: elasticity-based RTS using the IDF										
Sample I										
Total	1.0195	0.0527	0.8127	0.9699	0.9987	1.0119	1.0273	1.0928	1.5708	8,181
1	1.0302	0.0496	0.8194	0.9781	1.0077	1.0235	1.0432	1.1027	1.5214	2,041
2	1.0181	0.0404	0.8773	0.9786	1.0012	1.0131	1.0249	1.0706	1.5136	2,056
3	1.0083	0.0412	0.8127	0.9708	0.9954	1.0059	1.0159	1.0439	1.4854	2,060
4	1.0216	0.0716	0.8127	0.9537	0.9923	1.0089	1.0265	1.1445	1.5708	2,024
Sample II										
Total	1.0398	0.0449	0.8833	0.9897	1.0194	1.0340	1.0534	1.0981	1.5415	10,016
1	1.0563	0.0460	0.8833	1.0105	1.0333	1.0506	1.0688	1.1153	1.5045	2,521
2	1.0439	0.0388	0.8851	1.0099	1.0259	1.0372	1.0518	1.0937	1.5231	2,512
3	1.0303	0.0293	0.9034	1.0012	1.0166	1.0257	1.0388	1.0764	1.5088	2,518
4	1.0283	0.0556	0.8909	0.9570	1.0033	1.0234	1.0456	1.0999	1.5415	2,465
Sample III										
Total	1.0206	0.0943	0.7694	0.9536	0.9912	1.0060	1.0265	1.1127	2.2334	9,454
1	1.0278	0.0890	0.7769	0.9571	0.9980	1.0154	1.0385	1.1094	2.0604	2,373
2	1.0136	0.0675	0.7694	0.9586	0.9926	1.0062	1.0211	1.0782	2.1560	2,377
3	1.0080	0.0717	0.7715	0.9582	0.9899	1.0009	1.0138	1.0628	2.1596	2,378
4	1.0332	0.1332	0.7726	0.9382	0.9871	1.0044	1.0321	1.2032	2.2334	2,326
Panel B: elasticity-based RTS using the cost function										
Sample I: unadjusted RTS estimates										
Total	1.1731	0.1816	0.5613	0.9641	1.0975	1.1614	1.2220	1.3985	3.5603	8,181
1	1.1953	0.1834	0.5656	1.0369	1.1293	1.1718	1.2232	1.3989	3.5603	2,049
2	1.1848	0.1407	0.5786	1.0317	1.1213	1.1754	1.2302	1.3550	3.1433	2,062
3	1.1633	0.1313	0.5842	1.0021	1.0948	1.1581	1.2173	1.3537	2.3761	2,063
4	1.1486	0.2465	0.5613	0.8575	1.0268	1.1152	1.2158	1.5300	3.5067	2,007
Sample I: adjusted RTS estimates										
Total	1.0504	0.1112	0.5602	0.9231	1.0075	1.0401	1.0768	1.1988	2.2932	8,181
1	1.0744	0.1254	0.6164	0.9558	1.0228	1.0546	1.0970	1.2453	2.2932	2,045
2	1.0562	0.0852	0.5707	0.9786	1.0217	1.0487	1.0751	1.1725	2.1336	2,057
3	1.0413	0.0747	0.5602	0.9630	1.0091	1.0319	1.0641	1.1477	1.9407	2,060
4	1.0293	0.1413	0.5685	0.8552	0.9666	1.0154	1.0696	1.2385	2.2364	2,019

Notes: Panel A shows summary statistics for elasticity-based RTS for samples I, II, and III by size quartiles based on total assets. IDF denotes estimates obtained from the IDF using (16). Panel B does the same from the cost function using (15)

that not all the largest banking organization in the U.S. exhibit increasing returns to scale and that scale economies, when they are present, are small.

6 Conclusions

The debate over TBTF in the U.S., the U.K., and elsewhere renewed the interest in the study of economies of scale in the banking industry. If large banking organizations enjoy economies of scale that benefit society as a whole, breaking up the biggest banks into smaller institutions may have deleterious consequences for the economy and the development of the banking industry. Our work sheds light on the nature and existence of scale economies at the biggest banking organizations in the U.S.

We derive new measures of RTS and estimate them using nonparametric methods for large BHC and Commercial Banks (CB) in the U.S. from 2001 to 2010. In line with conventional wisdom, we find that not all BHC and CB exhibit IRTS. In addition, economies of scale for those banking organizations experiencing IRTS seem small. Thus, from the scale economies' viewpoint, breaking up large banking organizations into smaller institutions may impose little costs on the economy as a whole.

We believe that our methodology offers several advantages over recent studies in the literature. First, instead of using a cost function, we use an IDF which requires no information on input prices. Input price data are not readily available, and one needs to construct them based on expenditure and input data, thus adding additional noise to the estimation. Second, since the use of IDF does not require input price information, our model of bank production can include physical and financial capitals as regular inputs. Traditionally these variables are either excluded from the analysis or treated as quasi-fixed inputs in a cost function model given that their prices are not readily available [e.g., Wheelock and Wilson (2012) and Berger and Mester (2003)]. Third, our sample period is more relevant for the current policy debate regarding the existence of economies of scale at large banking organizations, since it covers a more recent time period. Fourth, our shorter sample period allows us to avoid using data dimension reduction techniques that may lead to noisier estimates of returns to scale, without sacrificing precision and interpretation of the results.

We also estimate RTS nonparametrically using the traditional elasticity-based approach. We find that our results are robust. In particular, both approaches give essentially the same results concerning the distribution and magnitude of RTS for the biggest banking institutions. Thus, we do find any empirical evidence to support the claims that elasticity-based nonparametric return to scale measures are inadequate.

Acknowledgments We thank Christopher F. Parmeter, Emir Malikov, James Byder, Jimmy Saravia, and three anonymous referees for their helpful comments which substantially improved the quality of the article. We, alone, are responsible for any remaining errors.

References

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

Caves DW, Christensen LR, Swanson JA (1981) Productivity growth, scale economies, and capacity utilization in U.S. railroads, 1955–74. Am Econ Rev 71(5):994–1002

Das A, Kumbhakar SC (2012) Productivity and efficiency dynamics in indian banking: an input distance function approach incorporating quality of inputs and outputs. J Appl Econom 27(2):205–234

Du P, Parmeter CF, Racine JS (2013) Nonparametric Kernel regression with multiple predictors and multiple shape constraints. Stat Sin 23(3):1343–1372
Färe R (1988) Fundamentals of production theory. Springer, Berlin
Färe R, Primont D (1995) Multi-output production and duality: theory and applications. Springer, New York
Feng G, Serletis A (2010) Efficiency, technical change, and returns to scale in large US banks: panel data evidence from an output distance function satisfying theoretical regularity. J Bank Financ 34(1):127–138
Feng G, Zhang X (2012) Productivity and efficiency at large and community banks in the US: a Bayesian true random effects stochastic distance frontier analysis. J Bank Financ 36(7):1883–1895
Haldane A (2012) On being the right size. Speech given at Institute of Economic Affairs' 22nd annual series, The 2012 Beesley lectures at the Institute of Directors, Pall Mall. URL: www.bankofengland.co.uk/publications/pages/speeches/2012/615.aspx. Accessed 13 Mar 2013
Hall P, Huang LS (2001) Nonparametric Kernel regression subject to monotonicity constraints. Ann Stat 29(3):624–647
Hall P, Li Q, Racine JS (2007) Nonparametric estimation of regression functions in the presence of irrelevant regressors. Rev Econ Stat 89(4):784–789
Härdle WK, Mammen E (1993) Comparing nonparametric versus parametric regression fits. Ann Stat 21(4):1926–1947
Harrison WB (2012) In defense of big banks. New York Times, URL: www.nytimes.com/2012/08/23/opinion/dont-break-up-the-big-banks.html. Accessed 15 Feb 2013
Hayfield T, Racine JS (2008) Nonparametric econometrics: the np package. J Stat Softw 27(5):1–32
Hughes JP, Mester LJ (2013) Who said large banks don't experience scale economies? Evidence from a risk-return-driven cost function. J Financ Intermed 22(4):559–585
Johnson S (2012) Tarullo telegraphs fed's plans to cap bank size. Bloomberg News' Column. URL: www.bloomberg.com/news/2012-12-09/tarullo-telegraphs-fed-s-plans-to-cap-bank-size.html. Accessed 27 Mar 2013
Li Q, Racine JS (2004) Cross-validated local linear nonparametric regression. Stat Sin 14(2):485–512
Racine JS, Li Q (2004) Nonparametric estimation of regression functions with both categorical and continuous data. J Econom 119(1):99–130
Restrepo-Tobón D, Kumbhakar SC, Sun K (2012) Are U.S. commercial banks too big? Tech. rep., Binghamton University, EAFIT University, and Aston University
Rosenblum H (2011) Choosing the road to prosperity: Why we must end too big to fail-now. Tech. rep., Federal Reserve Bank of Dallas, 2011 annual report
Sealey C Jr, Lindley JT (1977) Inputs, outputs, and a theory of production and cost at depository financial institutions. J Financ 32(4):1251–1266
Tarullo DK (2012a) Financial stability regulation. Speech at the distinguished jurist lecture, University of Pennsylvania Law School, Philadelphia. URL: www.federalreserve.gov/newsevents/speech/tarullo20121010a.htm. Accessed 26 Mar 2013
Tarullo DK (2012b) Industry structure and systemic risk regulation. Speech at the Brookings Institution Conference on structuring the financial industry to enhance economic growth and stability, Washington D.C. URL: www.federalreserve.gov/newsevents/speech/tarullo20121204a.htm. Accessed 26 Mar 2013
Tarullo DK (2012c) Regulation of foreign banking organizations. Speech at the Yale School of Management Leaders Forum, New Haven. URL: www.federalreserve.gov/newsevents/speech/tarullo20121128a.htm. Accessed 26 Mar 2013
Wack K (2012) Big-bank breakup popular with rank and file of both parties. URL: www.americanbanker.com/issues/177_172/big-bank-breakup-popular-with-rank-and-file-1052378-1.html. Accessed 28 Feb 2013
Wallison PJ (2012) Breaking up the big banks: Is anybody thinking? American Enterprise Institute for Public Policy Research. URL: www.aei.org/outlook/economics/financial-services/banking/breaking-up-the-big-banks-is-anybody-thinking/. Accessed 28 Feb 2013
Wang M, Van Ryzin J (1981) A class of smooth estimators for discrete distributions. Biometrika 68(1):301–309
Wheelock DC, Wilson PW (2011) Are credit unions too small? Rev Econ Stat 93(4):1343–1359
Wheelock DC, Wilson PW (2012) Do large banks have lower costs? New estimates of returns to scale for U.S. banks. J Money Credit Bank 44(1):171–199