WORKING PAPER SERIES

New Evidence on Returns to Scale and Product Mix Among U.S. Commercial Banks

David C. Wheelock
Paul W. Wilson

Working Paper 1997-003A
http://research.stlouisfed.org/wp/1997/97-003.pdf

PUBLISHED: Journal of Monetary Economics, June 2001

FEDERAL RESERVE BANK OF ST. LOUIS
Research Division
411 Locust Street
St. Louis, MO 63102

New Evidence on Returns to Scale and Product Mix Among U.S. Commercial Banks

February 1997

Abstract

Numerous studies have found that banks exhaust scale economies at low levels of output, but most are based on the estimation of parametric cost functions which misrepresent bank cost. Here we avoid specification error by using nonparametric kernel regression techniques. We modify measures of scale and product mix economies introduced by Berger *et al.* (1987) to accommodate the nonparametric estimation approach, and estimate robust confidence intervals to assess the statistical significance of returns to scale. We find that banks experience increasing returns to scale up to approximately $500 million of assets, and essentially constant returns thereafter. We also find that minimum efficient scale has increased since 1985.

JEL Classification: C14, E59, G21, L11

David C. Wheelock
Research Officer
Federal Reserve Bank of St. Louis
411 Locust Street
St. Louis, MO 63102

*Paul W. Wilson
Department of Economics
University of Texas
Austin, TX 78712

We are grateful to David Humphrey and participants of the 1997 Texas Econometerics Camp for comments on an earlier draft. Any remaining errors, of course, are our responsibility.

1. INTRODUCTION

The ongoing merger wave in the US banking industry has helped to eliminate nearly one-third of American commercial banks since 1984 (from 14,419 banks in 1984 to 9,919 at the end of 1995). The bulk of those eliminated have been small banks, and the disappearance of many small banks through acquisition and failure suggests they may not be viable in today's environment.[1]

On the other hand, as a group small banks have often been more profitable than their larger counterparts, and because researchers have typically found little evidence of significant economies of scale in banking, one might wonder whether the recent substantial increase in average (and median) bank size reflects a trend away from efficient resource allocation. Moreover, the evidence suggests that "megamergers" among large banks have not produced significant cost savings (e.g., Berger and Humphrey, 1992; Boyd and Graham, 1991), even though bankers themselves often argue that mergers improve their banks' operating efficiency or help them achieve economies of scale.[2]

One explanation for the relative decline of small banks is suggested by Berger and Humphrey (1991). They find that inefficiencies associated with operating off the best-practice frontier ("X-inefficiency") tend to dominate scale and scope inefficiencies in commercial banking, and that small banks suffer more X-inefficiency than larger banks (see also Wheelock and Wilson, 1996). Perhaps the X-inefficiencies of small banks outweigh the apparent inefficient scale of large banks, and thereby explain the ongoing decline of small banks.

Recently, however, conventional wisdom about the lack of scale economies in banking has been questioned. Much of the extant evidence on scale economies is based on estimation of translog cost functions (or other parametric forms). Although the translog function has some desirable properties, it has been shown to represent bank costs poorly,

[1] Between 1984 and 1995, the number of banks with less than $300 million of assets fell from 13,676 to 8860, while the number of banks with more than $300 million of assets increased from 739 to 1059.

[2] Akhavein, Berger and Humphrey (1996) find that mergers of large banks tend to enhance profit efficiency, however, because of revenue gains when merged banks adjust their mix of outputs toward higher-value assets, such as loans.

especially for banks near the small and large extremes of the size range of banks. The fact that estimates of efficient scale vary widely and appear to depend on whether banks of all sizes are included in the research sample, or just banks of a particular size range, is evidence that the translog is a misspecification. Two studies using a nonparametric specification of bank costs (McAllister and McManus, 1993; Mitchell and Onvural, 1996) suggest that banks experience increasing returns to scale at least up to $500 million of assets, and constant returns thereafter. By contrast, estimates based on estimation of a translog model suggest that scale economies are exhausted at about $100 million of assets (when banks of all sizes are included in the sample), with decreasing returns for larger banks (McAllister and McManus, 1993).[3]

The rapid pace of consolidation within the banking industry poses a challenge for regulators who must consider questions of competition and market service in the approval process. It also raises questions about the impact of technological and regulatory change on market structure in general (see, *e.g.,* Berger, Kashyap and Scalise, 1995). Our research investigates scale and scope economies for the banking industry, and how they may have changed over the past decade. We employ a nonparametric approach and, unlike previous studies, examine the universe of banks (except those with missing or unusable data) rather than a restricted sample. Furthermore, we refine the scale and scope measures suggested by Berger, Hanweck and Humphrey (1987) to estimate economies over the range of data and to accommodate a nonparametric approach. Finally, we provide robust confidence intervals to assess the statistical significance of our estimates of scale and scope economies.[4]

---

[3]Among the many studies of scale economies in banking based on estimation of translog cost functions are Berger *et al.* (1987), Berger and Humphrey (1991), Clark (1996), Gropper (1991), Hunter *et al.* (1990), Hunter and Timme (1994), and Jagtiani and Khanthavit (1996). The latter three studies are based on samples of banks with at least $1 billion of assets and all find evidence of scale economies for banks of up to $2 billion of assets. Clark (1996) is also based on large banks and finds scale economies for banks of less than $3 billion of assets. The remaining studies are based on samples drawn from small-size banks or banks of all sizes, and they find that scale economies are exhausted at considerably smaller asset sizes (*e.g.,* $100–$200 million). Humphrey (1990) surveys earlier studies of scale economies.

[4]McAllister and McManus (1993) and Mitchell and Onvural (1996) estimate scale economies for restricted samples. McAllister and McManus (1993) do not test the statistical significance of scale economies or consider whether scale economies changed over time. Mitchell and Onvural (1996) find that the industry cost function shifted between 1986 and 1990 and formally test the statistical significance of scale economies, though their test requires that the error term of the cost function be normally distributed. We provide

Section 2 presents our modification of the Berger *et al.* (1987) measures of scale and product mix economies. In Section 3 we describe our model of bank cost. Section 4 describes our estimation method, and Section 5 presents our empirical findings.

2. MEASURING RETURNS TO SCALE AND PRODUCT MIX

Consider a multiple-output cost function $C(\boldsymbol{y})$, where $\boldsymbol{y} = [\, y_1 \quad \ldots \quad y_q \,]'$ denotes a vector of outputs. Berger *et al.* (1987) note that a firm producing outputs $\boldsymbol{y}$ is competitively viable if the cost of producing $\boldsymbol{y}$ by that firm is no greater than the scale-adjusted cost of jointly producing output bundle $\boldsymbol{y}$ by any other set of firms. That is, for any and all output vectors $\boldsymbol{y}^\ell \geq 0$ and $\theta > 0$ such that $\sum_\ell \boldsymbol{y}^\ell = \theta \boldsymbol{y}$,

$$C(\boldsymbol{y}) \leq \theta^{-1} \sum_\ell C(\boldsymbol{y}^\ell), \tag{2.1}$$

where ℓ indexes specific output vectors which may be summed to form $\boldsymbol{y}$. Unfortunately, there are no simple necessary and sufficient conditions for competitive viability; a complete examination of the question for a given firm would require comparing the costs of hypothetical firms producing an infinite variety of output vectors.[5] Of course this is not feasible. The typical procedure in banking studies has been to compare hypothetical, representative firms producing output vectors at the sample means of outputs within various bank size classes (*e.g.*, Berger *et al.*, 1987; Clark, 1996; Mitchell and Onvural, 1996).

For illustration, consider two banks A and B producing two outputs in quantities $\boldsymbol{y}^a = [\, y_1^a \quad y_2^a \,]$ and $\boldsymbol{y}^b = [\, y_1^b \quad y_2^b \,]$, respectively, as shown in Figure 1 (which we have adapted from Berger *et al.*, 1987). Ray scale economies (RSCE) can be measured as the elasticity of cost along a ray $\theta \boldsymbol{y}$ emanating from the origin, so that the product mix is held constant:

$$\text{RSCE}(\boldsymbol{y}) \equiv \left. \frac{\partial \log C(\theta \boldsymbol{y})}{\partial \log \theta} \right|_{\theta=1} = \sum_j \frac{\partial \log C(\boldsymbol{y})}{\partial \log y_j}, \tag{2.2}$$

separate estimates of scale and product mix economies for the universe of banks for the years 1985, 1989 and 1994, and our tests of statistical significance require no restrictive assumptions about the error terms of the cost function.

[5]Note that we are ignoring demand-side considerations here and throughout.

where j indexes the different outputs. In terms of Figure 1, RSCE for firm A would be measured along the ray OA, while RSCE for firm B would be measured along the ray OB. As Berger *et al.* (1987) note, RSCE is the multiproduct analog of marginal cost divided by average cost on a ray from the origin, with RSCE$(<, =, >)1$ implying (increasing, constant, decreasing) returns to scale as output is expanded along the ray from the origin. A firm for which RSCE $\neq 1$ is not competitively viable; either a smaller or a larger firm could drive it from a competitive market.

The RSCE measure does not reflect variations in product mix among firms of different sizes. The output mix of banks tends to vary with size, so Berger *et al.* (1987) propose two alternative measures of returns that commingle scale and product mix economies. They define expansion path scale economies (EPSCE) as the elasticity of incremental cost with respect to incremental output along a nonradial ray such as the one emanating from point A in Figure 1 and passing through point B. Formally,

$$\text{EPSCE}(\boldsymbol{y}^a, \boldsymbol{y}^b) \equiv \left. \frac{\partial \log \left[C\left(\boldsymbol{y}^a + \theta(\boldsymbol{y}^b - \boldsymbol{y}^a)\right) - C\left(\boldsymbol{y}^a\right) \right]}{\partial \log \theta} \right|_{\theta=1}. \tag{2.3}$$

Conditional on firm A being competitively viable, firm B is viable if and only if EPSCE$(\boldsymbol{y}^a, \boldsymbol{y}^b) = 1$, indicating that as output is expanded from point A along the ray AB, constant returns to scale prevail at point B (note that the expression in (2.3) is evaluated at $\theta = 1$). If EPSCE$(\boldsymbol{y}^a, \boldsymbol{y}^b)(<, =, >)1$, then (increasing, constant, decreasing) returns to scale prevail at point B along the ray AB. Under increasing (decreasing) returns to scale a combination of larger (smaller) firms could drive firm B from the market.

As an alternative to EPSCE, Berger *et al.* (1987) also propose expansion path subadditivity (EPSUB), which they define as

$$\text{EPSUB}(\boldsymbol{y}^a, \boldsymbol{y}^b) \equiv \frac{C(\boldsymbol{y}^a) + C(\boldsymbol{y}^b - \boldsymbol{y}^a) - C(\boldsymbol{y}^b)}{C(\boldsymbol{y}^b)}. \tag{2.4}$$

The numerator term $C(\boldsymbol{y}^b - \boldsymbol{y}^a)$ in (2.4) gives the cost of firm D in Figure 1. Collectively, firms A and D produce the same output as firm B. If EPSUB$(\boldsymbol{y}^a, \boldsymbol{y}^b) < 0$, then firm B is not competitively viable; *i.e.,* two firms producing output vectors $\boldsymbol{y}^a$ and $\boldsymbol{y}^b - \boldsymbol{y}^a$ have lower

combined total cost than firm B, but collectively produce the same output. Alternatively, if EPSUB$(\boldsymbol{y}^a, \boldsymbol{y}^b) > 0$, then (smaller) firm A should adjust its output vector toward that of (larger) firm B. Although both EPSCE and EPSUB are composite measures of scale and scope economies, EPSUB is closer in spirit to a measure of scope economies than EPSCE because EPSUB compares the cost of production at a given firm B with the cost of producing an identical level of output in two separate firms with different output mixes. EPSCE, on the other hand, measures the incremental cost of incremental output along the expansion path between two different-sized firms.

The RSCE, EPSCE and EPSUB measures developed by Berger *et al.* (1987) are typically evaluated at specific points in the data space by replacing $C(\boldsymbol{y})$ and $\partial C(\boldsymbol{y})/\partial y_i$ with estimates $\widehat{C}(\boldsymbol{y})$ and $\partial \widehat{C}(\boldsymbol{y})/\partial y_i$, respectively, in (2.2)–(2.4). The RSCE and EPSCE measures require estimation of derivatives of the cost function, which is problematic unless the underlying cost function is parametrically specified. McAllister and McManus (1993) and Mitchell and Onvural (1996) have suggested, however, that the parametric functional form typically used in studies of banking costs, *i.e.,* the translog function, misrepresents bank costs. We also find evidence that the translog functional form misspecifies bank costs, and opt for nonparametric methods to estimate the cost function. This requires that we recharacterize the measures of ray scale economies and expansion path scale economies to avoid estimating derivatives of the cost function.[6]

First, define

$$\mathcal{S}(\theta|\boldsymbol{y}) \equiv \frac{C(\theta \boldsymbol{y})}{\theta C(\boldsymbol{y})}. \tag{2.5}$$

It is straightforward to show that

$$\frac{\partial \mathcal{S}(\theta|\boldsymbol{y})}{\partial \theta} \left(\begin{matrix}\leq \\ = \\ >\end{matrix}\right) 0 \Longleftrightarrow \text{RSCE}(\boldsymbol{y}) \left(\begin{matrix}\leq \\ = \\ >\end{matrix}\right) 1; \tag{2.6}$$

[6]We use kernel regression methods to estimate the bank cost function. While there are several criteria one might use to choose an appropriate smoothing parameter in estimating the cost function itself, from a practical viewpoint there are no useful criteria for choosing the smoothing parameter for estimating derivatives of the cost function. In general, estimation of derivatives requires more smoothing (*i.e.,* larger bandwidths in the context of kernel smoothing) than in estimation of the original function, but how much more is unknown in typical empirical settings. Moreover, this problem is not unique to kernel regression methods; *e.g.,* local polynomial regression methods, k-nearest neighbor methods, *etc.* involve similar issues. Further discussion of this problem is provided in Appendix A.

i.e., $\mathcal{S}(\theta|\boldsymbol{y})$ is decreasing (constant, increasing) in θ if returns to scale are increasing (constant, decreasing) at $\theta\boldsymbol{y}$ along the ray from the origin. By replacing $C(\cdot)$ with estimates $\widehat{C}(\cdot)$ in (2.5), and allowing θ to vary, one can plot $\widehat{\mathcal{S}}(\theta|\boldsymbol{y})$ as a function of θ and examine returns to scale over entire rays from the origin. Moreover, using bootstrap methods described in Appendix A, one can obtain confidence bands for the estimated curve $\widehat{\mathcal{S}}(\theta|\boldsymbol{y})$.

Similarly, for a given pair of output vectors $(\boldsymbol{y}^a, \boldsymbol{y}^b)$, we define

$$\mathcal{E}(\theta|\boldsymbol{y}^a, \boldsymbol{y}^b) \equiv \frac{C\left(\boldsymbol{y}^a + \theta(\boldsymbol{y}^b - \boldsymbol{y}^a)\right) - C\left(\boldsymbol{y}^a\right)}{\theta\left[C\left(\boldsymbol{y}^b\right) - C\left(\boldsymbol{y}^a\right)\right]}. \tag{2.7}$$

Straightforward algebra reveals that

$$\frac{\partial \mathcal{E}(\theta|\boldsymbol{y}^a, \boldsymbol{y}^b)}{\partial \theta} \left(\begin{matrix}\geq \\ = \\ \leq\end{matrix}\right) 0 \Longleftrightarrow \text{EPSCE}(\boldsymbol{y}^a, \boldsymbol{y}^b) \left(\begin{matrix}\geq \\ = \\ \leq\end{matrix}\right) 1; \tag{2.8}$$

i.e., $\mathcal{E}(\theta|\boldsymbol{y}^a, \boldsymbol{y}^b)$ increasing (constant, decreasing) in θ indicates decreasing (constant, increasing) returns to scale along the ray from $\boldsymbol{y}^a$ through $\boldsymbol{y}^b$ in the input/output space. In terms of Figure 1, consider values $0 < \theta \leq 1$. From (2.7), it is clear that $\mathcal{E}(\theta|\boldsymbol{y}^a, \boldsymbol{y}^b) = 1$ when $\theta = 1$. But if $\theta < 1$, then the first term in the numerator of (2.7) gives the cost of a hypothetical firm producing at an intermediate point along the segment AB. If $\theta < 1$ and $\mathcal{E}(\theta|\boldsymbol{y}^a, \boldsymbol{y}^b) > 1$, then the cost of this hypothetical firm is greater than the weighted costs of firms A and B, given by the numerator in (2.7). This implies that the cost surface is concave from below along the path AB, which in turn implies that total cost is increasing at a decreasing rate as we move from point A to point B in Figure 1. Hence if $\mathcal{E}(\theta|\boldsymbol{y}^a, \boldsymbol{y}^b) > 1$ for values $\theta < 1$, returns to scale are increasing along the expansion path AB. Note that we do not need to estimate the deriviative in (2.8) to draw this conclusion. Similar reasoning demonstrates that if $\mathcal{E}(\theta|\boldsymbol{y}^a, \boldsymbol{y}^b) < 1$ for values $\theta < 1$, decreasing returns to scale prevail along the expansion path AB.[7] As before, we replace $C(\cdot)$ with estimates $\widehat{C}(\cdot)$ in (2.7) to obtain estimates $\widehat{\mathcal{E}}(\theta|\boldsymbol{y}^a, \boldsymbol{y}^b)$, which can then be plotted as a function of θ to examine returns to scale along the entire ray from $\boldsymbol{y}^a$ through $\boldsymbol{y}^b$, rather than merely

[7] Conceivably, $\mathcal{E}(\theta|\boldsymbol{y}^a, \boldsymbol{y}^b)$ could oscillate around unity for values $\theta < 1$, which would suggest both increasing and decreasing returns along different parts of the expansion path.

at a single point along this ray. Also as before, we can use bootstrap methods to obtain confidence bands for the estimated curve $\widehat{\mathcal{E}}(\theta|\boldsymbol{y}^a, \boldsymbol{y}^b)$.

Finally, to measure EPSUB, we replace $\boldsymbol{y}^b$ in (2.4) with $\boldsymbol{y}^a + \theta(\boldsymbol{y}^b - \boldsymbol{y}^a)$ to obtain

$$\mathcal{A}(\theta|\boldsymbol{y}^a, \boldsymbol{y}^b) \equiv \frac{C(\boldsymbol{y}^a) + C\left(\theta(\boldsymbol{y}^b - \boldsymbol{y}^a)\right) - C\left(\boldsymbol{y}^a + \theta(\boldsymbol{y}^b - \boldsymbol{y}^a)\right)}{C\left(\boldsymbol{y}^a + \theta(\boldsymbol{y}^b - \boldsymbol{y}^a)\right)}. \tag{2.9}$$

Estimates $\widehat{\mathcal{A}}(\theta|\boldsymbol{y}^a, \boldsymbol{y}^b)$ may be obtained by replacing $C(\cdot)$ with $\widehat{C}(\cdot)$, and as with the other measures, $\widehat{\mathcal{A}}(\theta|\boldsymbol{y}^a, \boldsymbol{y}^b)$ can be plotted as a function of $0 < \theta \leq 1$ for given pairs $(\boldsymbol{y}^a, \boldsymbol{y}^b)$ to examine EPSUB along the entire path from A to B. The interpretation of $\mathcal{A}(\theta|\boldsymbol{y}^a, \boldsymbol{y}^b)$ is similar to the interpretation of the original measure in (2.4); *i.e.,* values greater than zero imply that the smaller firm should adjust its outputs toward those of the larger firm, while values less than zero imply that larger firms should be split into smaller firms, perhaps producing different output mixes. In terms of Figure 1, to the extent that $\boldsymbol{y}^a$ does not fall on the path $\theta\boldsymbol{y}^b$, values $\mathcal{A}(\theta|\boldsymbol{y}^a, \boldsymbol{y}^b) > 1$ provide some indication of economies of scope. Indeed, if the "small" firm were at point E in Figure 1, then $\mathcal{A}(\theta|\boldsymbol{y}^e, \boldsymbol{y}^b)$ would provide a measure of scope economies.

3. A MODEL OF BANK COST

To estimate the measures of scale and product mix economies described in the previous section, a model of bank cost must be specified. Banks use a number of inputs to produce a myriad of financial services, and in studies of bank technology researchers are forced to employ simplified models of bank production. Typically, banks are viewed as transforming various financial resources, as well as labor and physical plant, into loans, other investments and, sometimes, deposits. One view, termed the *production* approach, measures bank production in terms of the numbers of loans and deposit accounts serviced. The more common *intermediation* approach measures outputs in terms of the dollar amounts of loans and deposits. The production approach includes only operating costs, whereas the intermediation approach includes both operating costs and interest expense, and hence is probably of more interest for studying the viability of banks. We adopt the

the intermediation approach in this study.[8]

Researchers have used various criteria to identify the specific inputs and outputs to include in models of bank production. Typically, various categories of loans are treated as outputs, while funding sources, labor and physical plant are treated as inputs. The categorization of deposits varies across studies. Whereas non-transactions deposits are almost always treated as inputs, transactions deposits are sometimes considered to be outputs. Without a consensus on the specification of an input/output mapping, we follow Kaparakis *et al.* (1994), which is somewhat representative.

Data for this study are taken from the quarterly Statements of Income and Condition (call reports) filed by commercial banks. We use annual data for 1985, 1989, and 1994 to examine whether returns to scale and other aspects of bank costs have changed over recent history. Following Kaparakis *et al.* (1994), we specify four outputs, four variable inputs, and one quasi-fixed input for each bank $i = 1, \ldots, N$ in a given cross-sectional sample:[9]

Outputs:[10]

y_{i1} = loans to individuals for household, family, and other personal expenses;
y_{i2} = real estate loans;
y_{i3} = commercial and industrial loans;
y_{i4} = federal funds sold, securities purchased under agreements to resell, plus total securities held in trading accounts;

Variable inputs:

x_{i1} = interest-bearing deposits except certificates of deposit greater than $100,000;

[8]See Berger *et al.* (1987) or Ferrier and Lovell (1990) for further discussion of these approaches.

[9]Here and elsewhere, we denote the number of observations in a given cross section as N, although the number of observations varies across the three cross sections.

[10]Stocks used to define inputs and outputs (as opposed to flows used to define price variables) are mean values for the calendar year. For example, to compute outputs for 1985, we add the values of each stock from the end-of-year Call Reports for 1984 and 1985, and then divide by 2. All values are book values, except in the case of total securities held in trading accounts, which are reported in terms of market value beginning in 1994. Unfortunately, there are no periods for which both book and market values of these securities are available, which would allow direct comparison. However, we believe the effects of this data discrepancy are small since this item represents a small proportion of the fourth output for most banks. Moreover, we tried deleting this item in computing the fourth output, with only very minor effects on our quantitative results and no effect on the qualitative results. Our reported results include securities held in trading accounts in the fourth output.

x_{i2} = purchased funds (certificates of deposit greater than \$100,000, federal funds purchased, and securities sold plus demand notes) and other borrowed money;
x_{i3} = number of employees;
x_{i4} = book value of premises and fixed assets;

Quasi-fixed input:
x_{i5} = noninterest-bearing deposits.

Kaparakis *et al.* (1994) argue that since, by definition, banks cannot attract more noninterest-bearing deposits by offering interest, they should be regarded as exogenously determined as a first approximation. Although banks might offer various services or other incentives to attract non-interest bearing deposits, we assume that banks take the quantity of these deposits as given.[11] Because no explicit price exists for this input, it must either be omitted from the cost function altogether, or its quantity rather than price must be included in the cost function. Like Kaparakis *et al.*, we opt for the latter approach.

Input prices are computed as follows:

Input prices:
p_{i1} = average interest cost per dollar of x_{i1};
p_{i2} = average interest cost per dollar of x_{i2};
p_{i3} = average wage per employee;
p_{i4} = average cost of premises and fixed assets.

Finally, the total cost of the variable inputs defines the dependent variable C_i to be used in estimating the cost function; *i.e.*,

$$C_i = \frac{1}{p_{i4}} \sum_{j=1}^{4} p_{ij} x_{ij}. \tag{3.1}$$

Costs are normalized by p_{i4} in (3.1) to ensure linear homogeneity of costs in input prices.

[11] With the recent proliferation of banks offering "sweep" accounts in which noninterest-bearing transactions deposits are automatically moved into interest-earning accounts until needed, the treatment of noninterest-bearing accounts as a quasi-fixed input may be less tenable. Prior to 1995, however, few banks offered such accounts and so our treatment of noninterest bearing accounts as quasi-fixed seems reasonable. See Hunter and Timme (1995) for further discussion of quasi-fixed inputs and an investigation of the empirical significance of taking core deposits as a quasi-fixed input for a sample of large banks.

Because the Call Reports include some firms that have bank charters, but which do not function as traditional banks (*e.g.*, credit-card subsidiaries of bank holding companies), we employ several selection criteria to limit the sample to a group of relatively homogenous banks. In particular, we omit banks reporting negative values for inputs, outputs, or prices. We convert all dollar values to 1992 prices using the GDP deflator. Since some remaining observations contain values for p_{i1} and p_{i2} that are suspect, we omit observations when either of these variables exceed 0.25.[12] After omitting observations with missing or implausible values, we have 13,168, 11,786, and 9,819 observations for 1985, 1989, and 1994, respectively.

The decreasing numbers of observations in our three samples is consistent with the decline in the number of US commercial banks by about one-third between 1985 and 1994. During this same period, the mean (and median) bank size increased. The relatively large decline in the number of small banks suggests that such banks became less competitively viable over the period. Figure 2 shows kernel estimates of the densities of the log of total assets for each year in our study. Kolmogorov-Smirnov two-sample tests of the null hypothesis of no difference in the distributions across time are rejected at 99.5 percent significance for each pair of years 1985/1989, 1989/1994, and 1985/1994.[13] Because the distribution of bank sizes shifted over time, we investigate whether returns to scale changed similarly over time.[14]

4. ESTIMATION METHOD

Having specified outputs and input prices, we must estimate the relation between these variables and bank costs in order to estimate the RSCE, EPSCE and EPSUB measures discussed above. The data described in the previous section may be represented by

[12] We arrived at this criteria by examining the distributions of the price variables; the distributions were somewhat continuous up to some point below 0.25, with a few (clearly implausible) large outliers in the right tail.

[13] The Kolmogorov-Smirnov test is unaffected by taking the log of total assets. The kernel density estimates were obtained using a standard Gaussian kernel function; optimal bandwidths were approximated for each year using the least-squares cross-validation procedure described by Silverman (1986).

[14] The rightward shift of the distribution of bank sizes over time conceivably reflects changes in the underlying technology, calling into question the approach used by McAlister and McManus (1993), where annual observations are pooled over time.

the partitioned matrix $[\boldsymbol{X} \quad \boldsymbol{C}]$, where $\boldsymbol{C}$ is an $(N \times 1)$ column vector whose elements C_i represent normalized variable costs for banks $i = 1, \ldots, N$, and $\boldsymbol{X}$ is an $(N \times K)$ matrix of explanatory variables, with the ith row of $\boldsymbol{X}$ equal to

$$\boldsymbol{X}_i = [X_{ij}] = [\, y_{i1} \quad y_{i2} \quad y_{i3} \quad y_{i4} \quad p_{i1}/p_{i4} \quad p_{i2}/p_{i4} \quad p_{i3}/p_{i4} \quad x_{i5} \,] \tag{4.1}$$

(hence for the present application, given the input/output mapping outlined in the previous section, $K = 8$). Costs and input prices are normalized with respect to the fourth input price to ensure homogeneity of the cost function with respect to input prices.

In order to infer scale efficiencies among banks, the mapping $\boldsymbol{C} \leftarrow \boldsymbol{X}$ must be estimated. The usual approach involves estimating the conditional expectation function $m(\boldsymbol{x}) = E(C_i | \boldsymbol{X}_i = \boldsymbol{x})$. Assume

$$C_i = m(\boldsymbol{X}_i) + \varepsilon_i, \qquad i = 1, \ldots, N, \tag{4.2}$$

where ε_i is an independent stochastic error term, $E(\varepsilon_i | \boldsymbol{X}_i = \boldsymbol{x}) = 0$, and $\mathrm{VAR}(\varepsilon_i | \boldsymbol{X}_i = \boldsymbol{x}) = \sigma^2(\boldsymbol{x})$. In addition, assume the observations $(\boldsymbol{X}_i, C_i)$ are multivariate identically and independently distributed across i.

Typical studies of bank costs have used parametric specifications for the conditional mean function; by far, the most common choice of functional forms has been the translog specification. For example, Kaparakis *et al.* (1994) condition on $\boldsymbol{X}$ and use a translog specification equivalent to

$$\begin{aligned} \log C_i = {} & [\, 1 \quad \boldsymbol{A} \,] \boldsymbol{\alpha} + \boldsymbol{B}\boldsymbol{\beta} + \boldsymbol{A}\boldsymbol{\Pi}\boldsymbol{A}' + \boldsymbol{B}\boldsymbol{\Delta}\boldsymbol{B}' + \boldsymbol{A}\boldsymbol{\Gamma}\boldsymbol{B}' \\ & + \boldsymbol{\Phi} [\, 1 \quad \boldsymbol{A} \,]' (\log x_{i5}) + \boldsymbol{\Psi}\boldsymbol{B}' + \tau (\log x_{i5})^2 + \epsilon_i, \end{aligned} \tag{4.3}$$

where $\boldsymbol{A} = [\, \log y_{i1} \quad \ldots \quad \log y_{i4} \,]$, $\boldsymbol{B} = [\, \log(p_{i1}/p_{i4}) \quad \log(p_{i2}/p_{i4}) \quad \log(p_{i3}/p_{i4}) \,]$; $\boldsymbol{\alpha} = [\, \alpha_0 \quad \ldots \quad \alpha_4 \,]'$, $\boldsymbol{\beta} = [\, \beta_1 \quad \beta_2 \quad \beta_3 \,]'$, $\boldsymbol{\Pi} = [\, \pi_{jk} \,]$ with dimensions (4×4), $\boldsymbol{\Delta} = [\, \delta_{jk} \,]$ with dimensions (3×3), $\boldsymbol{\Gamma} = [\, \gamma_{jk} \,]$ with dimensions (4×3), $\boldsymbol{\Phi} = [\, \phi_0 \quad \ldots \quad \phi_4 \,]$, and $\boldsymbol{\Psi} = [\, \psi_1 \quad \phi_2 \quad \phi_3 \,]$ are arrays of parameters to be estimated; τ is a scalar parameter to be estimated; and $\pi_{jk} = 0 \; \forall \; j > k$, $\delta_{jk} = 0; \forall \; j > k$.[15] Aside from the problem of

[15] Cost inefficiency is usually measured using a composite error term. To the extent that cost inefficiency is present here, the error term in (4.3) might be skewed.

having to delete or modify observations with zero values in order to take logs, the translog specification is flexible only in a local sense and, as demonstrated below, misspecifies bank costs over the observed range of bank sizes.

We formally test the translog specification of bank cost in (4.3) using our data; details are given in Appendix B. As McAllister and McManus (1993) note, the translog cost function was originally developed as a local approximation to some unknown "true" underlying cost function. This raises suspicion about the translog's ability to replicate the true cost function in banking studies, where data are typically highly dispersed. Indeed, our data lead us to reject the translog specification at any reasonable level of significance. Some authors have implicitly recognized this problem and have restricted their studies to samples of banks from within a narrow size range. This seems an odd approach for examining scale economies or finding the most efficient scale in the banking industry and, as McAllister and McManus note, it is likely to generate misleading results.

Rejection of the translog functional form points to the use of nonparametric estimation methods. Although nonparametric methods are less efficient in a statistical sense than parametric methods when the true functional form is known, nonparametric estimation does not incur the risk of specification error. Moreover, the regression technique we use below does not require deleting or adjusting output observations with zero values, as required with the translog function.

An intuitively appealing way to estimate the conditional mean function $m(\boldsymbol{x})$ without imposing functional forms *a priori* is to use kernel methods to first compute an estimate of the joint density $f(\boldsymbol{x}, c)$ of $(\boldsymbol{X}_i, C_i)$ and then to integrate to obtain an estimate of

$$m(\boldsymbol{x}) = \frac{\int c f(\boldsymbol{x}, c)\, dc}{\int f(\boldsymbol{x}, c)\, dc}. \tag{4.4}$$

Substituting kernel density estimates for $f(\boldsymbol{x}, c)$ in the numerator and denominator of (4.4) yields the familiar Nadarya-Watson estimator of the conditional mean function (Nadarya, 1964; Watson, 1964).

It is well-known that kernel estimators such as the Nadarya-Watson estimator suffer from the curse of dimensionality; *i.e.,* for a given sample size, mean square error increases

dramatically with the dimensionality of the sample space. Silverman (1980) illustrates the problem in the context of density estimation by giving the sample sizes required to ensure that the relative mean square error of the kernel estimate of a standard multivariate Gaussian density at zero is less than 0.1, using a Gaussian kernel and the bandwidth that minimizes mean square error at zero, for various dimensions of the sample space. The required sample sizes reported by Silverman are 19, 223, 2790, and 43700 for 2, 4, 6, and 8 dimensions, respectively. Recall that $\boldsymbol{X}_i$ is (1×8) as specified in (4.1) above; with sample sizes ranging from 9,819 to 13,168 in this study, it would appear that direct estimation of the conditional mean function using the $\boldsymbol{X}_i$ would indeed incur the curse of dimensionality.

To deal with this problem, we use a data reduction method based on principal components transformations suggested by Scott (1992); details are given in Appendix A. For each of the cross-sections we examine, we are able to reduce the dimensionality from 8 to 5, which, given our sample sizes, should give reasonably accurate estimates of the cost function.

5. EMPIRICAL ANALYSIS

To implement the kernel regression methods outlined above and in Appendix A, we must first choose appropriate values for the bandwidth parameter, h. Using the full samples for 1985, 1989, and 1994, minimizing the least-squares crossvalidation function defined in equation (A.19) of Appendix A yields optimal bandwidths of 0.2601, 0.2306, and 0.2622 (corresponding to 1985, 1989, and 1994, respectively). Our method of transforming the data described in Appendix A enables us to use a single bandwidth parameter in estimating (4.4), rather than a vector or matrix of such parameters. Some experimentation with alternative values suggested that our qualitative results are not very sensitive to small changes in the bandwidths.

Using the methods described in Appendix A, we compute estimates $\widehat{C}$ of bank costs C, and then substitute these estimates into (2.5) to obtain estimates $\widehat{\mathcal{S}}(\theta|\boldsymbol{y})$ of our RSCE measure. In computing $\widehat{\mathcal{S}}(\theta|\boldsymbol{y})$, we set each element of the vector $\boldsymbol{y}$ equal to the mean of the corresponding output over the entire sample; other arguments in

the cost function (namely, the normalized input prices and the quasi-fixed input) are set equal to their sample means as well. $\widehat{\mathcal{S}}(\theta|\boldsymbol{y})$ is then computed for specific values of θ: 0.05, 0.10, 0.15, ..., 0.95, 1, 2, ..., 148. In addition, we use the bootstrap procedure discussed in Appendix A to estimate simultaneous confidence intervals for $\mathcal{S}(\theta|\boldsymbol{y})$ at each of the above values of θ.[16]

The estimated values $\widehat{\mathcal{S}}(\theta|\boldsymbol{y})$ are plotted in Figure 3 for each year 1985, 1989, and 1994, with $\log(\theta)$ on the horizontal axis. The vertical bars denote estimated 95 percent simultaneous confidence intervals at values of θ corresponding to $\log\theta = -3.5, -2.5, \ldots, 4.5$. By definition, $\widehat{\mathcal{S}}(\theta|\boldsymbol{y}) = 1$ at $\theta = 1$ ($\log(\theta) = 0$). The confidence intervals become smaller as $\log\theta$ rises, and are barely visible at $\log\theta = 4.5$ with the scaling in our figures. One might expect confidence intervals to widen moving away from the center of the data (corresponding to $\log\theta = 0$ here). However, $\mathcal{S}(\theta)$ is not a conditional mean function, but rather involves the ratio of two conditional mean functions. There is likely to be substantial correlation between the numerator and denominator in $\widehat{\mathcal{S}}(\theta)$, and this serves to reduce the width of the estimated confidence intervals for larger values of θ.

For reference, $\theta = 1$ ($\log(\theta) = 0$) corresponds to banks producing the mean output vector; presumably these correspond to banks with near the sample mean value of assets. In 1994, mean assets were \$152.8 million, and assets ranged from \$1.8 million to \$21.6 billion. Dividing these minimum and maximum values by mean assets we find that $\theta \approx 0.012$ ($\log\theta \approx -4.42$) for the smallest bank, and $\theta \approx 141.7$ ($\log\theta \approx 3.73$) corresponds to the largest bank.

The results in Figure 3 indicate that in each year, $\widehat{\mathcal{S}}(\theta|\boldsymbol{y})$ is mostly decreasing in θ, implying increasing returns to scale as discussed in Section 2. From left to right in Figure 3, $\widehat{\mathcal{S}}(\theta|\boldsymbol{y})$ is initially sharply downward sloping as output rises toward the mean bank size (corresponding to $\log(\theta) = 0$, or $\theta = 1$), implying dramatically increasing returns to scale for banks smaller than the mean size. Farther to the right the slope decreases, suggesting that while larger banks may also face increasing returns to scale, they are less

[16] We use a discrete number of simultaneous confidence intervals rather than confidence bands because they are easier to compute and to interpret, and impart almost as much information.

dramatic than for smaller banks. For 1985 and 1989, we can reject the null hypothesis of constant returns to scale in favor of increasing returns to scale between all successive pairs of locations of the confidence intervals. For 1994, we can do the same except between the last two confidence intervals, where a line with zero slope can pass through both intervals. Indeed, for 1994, $\widehat{S}(\theta|\boldsymbol{y})$ begins to increase with θ for values of θ between 36 and 50 ($3.5835 \leq \log(\theta) \leq 3.9120$, or asset-sizes of \$5.5 billion to \$7.6 billion), but then $\widehat{S}(\theta|\boldsymbol{y})$ again decreases in θ beyond this range. The simultaneous confidence intervals estimated in this region of the curve are quite narrow; we varied their location, each time finding similarly narrow confidence intervals, suggesting that the increase in $\widehat{S}(\theta|\boldsymbol{y})$ through the range $3.5835 \leq \log(\theta) \leq 3.9120$ is statistically significant. However, given the sparseness of banks in this size range, we are reluctant to make too much of this result.[17]

We next compute estimates $\widehat{\mathcal{E}}(\theta|\boldsymbol{y}^a, \boldsymbol{y}^b)$ of the EPSCE measure for $\theta = 0.1, 0.2, \ldots, 0.9$ by replacing C with kernel estimates $\widehat{C}$ of the cost function in (2.7). As before, we use the bootstrap methods discussed in Appendix A to estimate confidence intervals for $\mathcal{E}(\theta|\boldsymbol{y}^a, \boldsymbol{y}^b)$, although here we use pointwise rather than simultaneous confidence intervals since our interest lies only in whether $\widehat{\mathcal{E}}(\theta|\boldsymbol{y}^a, \boldsymbol{y}^b)$ is significantly different from unity at various values of θ, rather than in the slope of $\widehat{\mathcal{E}}(\theta|\boldsymbol{y}^a, \boldsymbol{y}^b)$. For $\boldsymbol{y}^a$ and $\boldsymbol{y}^b$, we use mean output vectors for the nine asset-size groups analyzed by Berger *et al.* (1987), with an additional group for banks with assets of greater than \$1 billion.[18]

These results are illustrated in Figure 4, where the vertical bars represent the estimated 95 percent confidence intervals, and the estimates $\widehat{\mathcal{E}}(\theta|\boldsymbol{y}^a, \boldsymbol{y}^b)$ for successive values

[17]Our results for RSCE are similar to those of McAllister and McManus (1993). Using kernel regression, they find increasing returns for banks with less than \$500 million of assets and roughly constant returns for larger banks (although in the absence of formal hypothesis testing, it is difficult to judge returns to scale with confidence). Using other nonparametic techniques (fourier transforms and spline functions), they find that returns to scale may be increasing for banks up to \$5 billion of assets (though again, formal hypothesis tests were not conducted). For additional comparison, we computed $\widehat{S}(\theta|\boldsymbol{y})$ using parameter estimates from the (misspecified) Translog cost funciton (4.3) for each year. In each instance, this produced a U-shaped curve for $\widehat{S}(\theta|\boldsymbol{y})$, with minimum values for 1985, 1989, and 1994 at \$3.270 billion, \$750.6 million, and \$1.375 billion of assets, respectively. For 1985, $\widehat{S}(\theta|\boldsymbol{y})$ based on the translog estimates increases slowly after reaching the minimum; but for 1989 and 1994, the increase is rather sharp. These differences may merely reflect the misspecification of bank costs.

[18]Other arguments of the cost function were evaluated at sample means.

of θ are joined by line segments. Figure 4 contains 18 graphs arranged in nine rows and three columns; the columns correspond to the years 1985, 1989, and 1994, while the rows correspond to successive pairs of asset-size categories as indicated by the pairs of ranges on the left. Vertical scales for graphs in a given row are the same, but differ across rows.

By definition, $\widehat{\mathcal{E}}(\theta|\boldsymbol{y}^a, \boldsymbol{y}^b) = 1$ when $\theta = 1$, and so the error bars in Figure 4 collapse as $\theta \to 1$. The estimated values $\widehat{\mathcal{E}}(\theta|\boldsymbol{y}^a, \boldsymbol{y}^b)$ exceed unity in most cases, indicating increasing returns to scale as banks expand their output vectors from the mean of the smaller size group to the mean of the larger size group. However, the estimated values are not significantly different from unity in a number of cases. For each pair of asset-size categories, the pattern of results are similar across the three years considered, although the statistical significance varies. In 1994 (third column of graphs in Figure 4), for the first pair of asset-size groups (\$0–10 million/\$10–25 million), $\widehat{\mathcal{E}}(\theta|\boldsymbol{y}^a, \boldsymbol{y}^b)$ is significantly different from 1 for each value of θ except $\theta = 0.1,\ 0.2,\ 0.3$. Thus, there are eventually increasing returns to scale along the path between mean outputs of these size groups. For the next two size groups, $\widehat{\mathcal{E}}(\theta|\boldsymbol{y}^a, \boldsymbol{y}^b)$ is significantly different from 1 for each value of θ considered, and for the \$50–75 million/\$75–100 million group comparison, $\widehat{\mathcal{E}}(\theta|\boldsymbol{y}^a, \boldsymbol{y}^b)$ is significantly different from 1 for $\theta = 0.1,\ 0.2$, but insignificantly different for larger values of θ. Thus while returns to scale appear to be initially increasing along this path, we cannot reject constant returns along the remainder of the path.

Similarly, we are unable to reject constant returns to scale for $\theta = 0.1, \ldots, 0.5$ over the expansion path from the mean of the \$75–100 million range to the mean of the \$100–200 million range, though we can reject constant returns for larger values of θ. We find increasing returns throughout, however, over the range from \$100–200 million to \$200–300 million, and from \$200–300 million to \$300–500 million. For the \$300–500 million/\$500–1000 million group comparison, $\widehat{\mathcal{E}}(\theta|\boldsymbol{y}^a, \boldsymbol{y}^b)$ is less than unity, suggesting decreasing returns to scale, but insignificantly so for $\theta = 0.1, \ldots, 0.4$ and the difference is only marginally significant for larger values of θ. As with Figure 3, caution must be used in interpreting results for the larger banks due to sparseness of the data in this region. Finally, for

the comparison of the largest size groups, we again find statistically significant increasing returns to scale along the expansion path.

For 1994, our RSCE and EPSCE measures of scale economies paint similar pictures of returns to scale in banking. We find evidence of increasing returns to scale for banks of at least $500 million of assets, and possibly much larger. Moreover, both measures suggest that returns to scale do not simply vary from increasing to constant to decreasing as size increases; rather, as size increases, there appear to be regions of increasing returns followed by constant returns, and then increasing returns again. For very large banks, there appears to be a region of decreasing returns surrounded by regions of constant returns, although the limited number of observations in this region cautions against drawing a firm conclusion.

For 1985 and 1989, the results for RSCE and EPSCE are again similar, though the RSCE measure indicates increasing returns to scale over a larger range of banks than does EPSCE. For 1985, we find little evidence of increasing returns for banks above the $200–300 million asset category using the EPSCE measure, but for 1989 the results suggest increasing returns for banks on the order of $500 million of assets. As for 1994, for 1985 there is some evidence of increasing returns in comparing the $500–1000 million asset size banks with those with assets in excess of $1 billion.

As with the RSCE and EPSCE measures, we computed estimates $\widehat{\mathcal{A}}(\theta|\boldsymbol{y}^a, \boldsymbol{y}^b)$ of the EPSUB measure in (2.9) for $\theta = 0.1,\ 0.2,\ \ldots,\ 1.0$ by replacing C with kernel estimates of the cost function. We again estimated 95 percent pointwise confidence intervals using the bootstap procedure described in Appendix A. Values of $\boldsymbol{y}^a$ and $\boldsymbol{y}^b$ were chosen as in the case of the EPSCE measure discussed above. These results are displayed in Figure 5, which is arranged similarly to Figure 4. In every case except in comparing the two largest size groups in 1989 (represented by the last graph in the middle column of Figure 5), we find that $\widehat{\mathcal{A}}(\theta|\boldsymbol{y}^a, \boldsymbol{y}^b)$ is significantly greater than 0, indicating that hypothetical banks producing the mean of the outputs of the smaller size group should expand their outputs. The EPSUB results thus broadly support those for RSCE and EPSCE, suggesting efficiencies for banks from increasing the size and, possibly, scope of their operations.

6. SUMMARY AND CONCLUSIONS

Conventional wisdom holds that banks exhaust scale economies at roughly $100–200 million of assets, approximately the mean bank size in 1985. This belief, however, is based on estimates of parametrically-specified cost functions for generally small samples of commercial banks. Such cost functions, including the translog function, have recently been shown to misrepresent bank costs, especially for banks near the extreme ranges of bank sizes—a point we verify here. In this paper, we present new evidence of scale and product mix economies based on kernel regression estimates of a nonparametric model of bank cost for the universe of commercial banks with usable data for 1985, 1989 and 1994. Our results are thus not subject to misspecification of the bank cost function nor to a limited sample.

Our estimates of scale and product mix economies are based on measures proposed by Berger *et al.* (1987). The rejection of the translog and other parametric cost function forms, however, necessitates modifications of the scale and product mix economy measures as derived here. Moreover, unlike previous studies, we present statistical tests of scale economies which do not assume that the errors of the cost function are normally distributed.

The results of this and other recent research based on nonparametric cost functions suggest that banks experience increasing returns to scale as they grow to approximately $500 million of assets or even larger. We find some limited evidence of decreasing returns to scale for banks of roughly $5.5 billion to $9 billion of assets (based on RSCE), but generally large banks appear to operate under constant returns over a wide range of sizes. Finally, at least one measure of scale economies—expansion-path scale economies—suggests that the bank size at which economies of scale are exhausted has increased since 1985. Our results are thus consistent with the ongoing increase in mean (and median) bank size, but also suggest that banks of considerably different sizes (though still large by historical standards) are competitively viable.

APPENDIX A

Regression Techniques with Dimension Reduction:

To estimate the conditional mean function in (4.4), we must first replace the joint density $f(\boldsymbol{x}, c)$ with an appropriate estimator, then perform the integration in the numerator and denominator. Provided the data have been transformed so that the elements of $\boldsymbol{x}$ and c are approximately identically and independently distributed, the joint density $f(\boldsymbol{x}, c)$ can be estimated by the kernel density estimator

$$\widehat{f}_h(\boldsymbol{x}, c) = N^{-1} \sum_{i=1}^{N} \underline{\mathcal{K}}_h(\boldsymbol{x} - \boldsymbol{X}_i)\mathcal{K}_h(c - C_i), \tag{A.1}$$

where

$$\mathcal{K}_h(\cdot) = h^{-1}\mathcal{K}(\cdot/h), \tag{A.2}$$

and $\underline{\mathcal{K}}_h(\cdot)$ is a multidimensional product kernel density estimator defined as

$$\underline{\mathcal{K}}_h(\cdot) = \prod_{j=1}^{d} \mathcal{K}_h(\cdot), \tag{A.3}$$

d is the length of the vector-valued argument to $\underline{\mathcal{K}}_h(\cdot)$, $\mathcal{K}(\cdot)$ is a kernel function, and h is the bandwidth which determines the extent to which the kernel estimator smoothes the empirical density function.[19] For consistent estimation, kernel functions must be piecewise continuous, symmetric about zero, and must integrate to unity; *i.e.*, $\mathcal{K}(t) = \mathcal{K}(-t)$ and $\int \mathcal{K}(t)\, dt = 1$. For purposes of this paper, we choose the standard Gaussian density as the kernel $\mathcal{K}(\cdot)$.[20]

---

[19] Choice of reasonable values for h needed to implement the kernel estimation method will be discussed later. The data can be transformed to be approximately iid by rescaling the data to have constant variance and zero covariance, and then suitably transforming each variable so that marginal distributions are similar. This allows use of a single bandwidth parameter. The actual rescaling and transformation of our data will be discussed in detail below.

[20] See Silverman (1986) for discussion of the merits of Gaussian kernels versus other choices such as Epanechnikov or quartic kernels. Gasser and Müller (1979, 1984), Müller (1988), and others have advocated use of high-order kernels for density estimation, typically in the form of even-order (and hence symmetric) polynomials over an interval such as $[-1, 1]$. While these kernels may reduce bias and may have faster rates of convergence, polynomial kernels produce negative regression weights, unlike when proper density functions are used as kernels. While preference for nonnegative regression weights is partly a matter of taste, Härdle and Carroll (1990) observe that the choice is not entirely idiosyncratic. In the context of the conditional mean function, it is difficult to find an intuitive interpretation of negative regression weights, regardless of the analytical niceties they might offer.

From the properties of kernel functions listed above, it is clear that

$$\int \widehat{f}(\boldsymbol{x}, c)\, dc = N^{-1} \sum_{i=1}^{N} \underline{\mathcal{K}}_h(\boldsymbol{x} - \boldsymbol{X}_i) \tag{A.4}$$

and

$$\int c\widehat{f}(\boldsymbol{x}, c)\, dc = N^{-1} \sum_{i=1}^{N} \underline{\mathcal{K}}_h(\boldsymbol{x} - \boldsymbol{X}_i) C_i. \tag{A.5}$$

Substituting these experessions into the numerator and denominator of (4.4) yields the Nadarya-Watson estimator of the conditional mean function:

$$\widehat{m}_h(\boldsymbol{x}) = \frac{\sum_{i=1}^{N} \underline{\mathcal{K}}_h(\boldsymbol{x} - \boldsymbol{X}_i) C_i}{\sum_{i=1}^{N} \underline{\mathcal{K}}_h(\boldsymbol{x} - \boldsymbol{X}_i)} \tag{A.6}$$

(Nadarya, 1964; Watson, 1964). This estimator has been discussed in detail by Müller (1988), Härdle (1990), and others.

We employ a principal components transformation suggested by Scott (1992) to reduce the mean square error of our estimates, which can be excessive with kernel estimation in high dimensional spaces. First, we use marginal transformations on the $K+1$ columns of the data matrix $[\,\boldsymbol{X} \quad \boldsymbol{C}\,]$ to construct $[\,\boldsymbol{X}^* \quad \boldsymbol{C}^*\,]$ such that elements within each of the columns of $[\,\boldsymbol{X}^* \quad \boldsymbol{C}^*\,]$ are approximately normally distributed.[21] Next, eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_K$ along with the corresponding $(K \times 1)$ eigenvectors $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_K$ of the sample correlation matrix

$$\widehat{\boldsymbol{R}}_K = (\operatorname{diag} \widehat{\boldsymbol{\Sigma}}_K)^{-1/2} \widehat{\boldsymbol{\Sigma}}_K (\operatorname{diag} \widehat{\boldsymbol{\Sigma}}_K)^{-1/2} \tag{A.7}$$

are computed, where

$$\widehat{\boldsymbol{\Sigma}}_K = N^{-1}(\boldsymbol{X}^* - \overrightarrow{\mathbf{1}}\, \overline{\boldsymbol{X}}^*)'(\boldsymbol{X}^* - \overrightarrow{\mathbf{1}}\, \overline{\boldsymbol{X}}^*) \tag{A.8}$$

is the sample covariance matrix, $\overrightarrow{\mathbf{1}}$ is an $(N \times 1)$ vector of ones, $\overline{\boldsymbol{X}}^* = N^{-1}\, \overrightarrow{\mathbf{1}}'\, \boldsymbol{X}^*$, and $\operatorname{diag} \widehat{\boldsymbol{\Sigma}}_K$ is a $K \times K$ matrix whose principal diagonal corresponds to that of $\widehat{\boldsymbol{\Sigma}}_K$ and whose off-diagonal elements are zero.

[21] This was accomplished by setting $C_i^* = \log(C_i)$, $X_{ij}^* = \log(X_{ij} + 0.01)\ \forall\ j = 1, \ldots, 4$ and $X_{ij}^* = \log(X_{ij})\ \forall\ j = 5, \ldots, 8$. The constant 0.01 was used for the output quantities $(j = 1, \ldots, 4)$ since these variables are observed at zero for some banks; using the constant avoids having to delete these observations, and casual examination of histograms of the resulting transformed variables suggest that the columns of $\boldsymbol{X}^*$ are approximately normally distributed.

The principal components transformation amounts to setting

$$\boldsymbol{T} = (\boldsymbol{X}^* - \overrightarrow{\mathbf{1}}\,\overline{\boldsymbol{X}}^*)\boldsymbol{E}, \tag{A.9}$$

where $\boldsymbol{E} = [\,\boldsymbol{e}_1, \ldots, \boldsymbol{e}_K\,]$. It is well-known that the principal components, *i.e.* the columns of $\boldsymbol{T}$, are uncorrelated, and $\mathrm{tr}\widehat{\boldsymbol{R}}_K = \sum_{j=1}^{K} \widehat{\lambda}_j = K$. It is equally well-known that there is no other linear combination of the columns of $\boldsymbol{X}^*$ with larger variance than the first principal component (*i.e.*, the first column of $\boldsymbol{T}$), and that the second principal component is the linear combination with the second-largest variance, *etc.*

Since the goal is to remove dimensions that contain no independent linear information, the reduced dimension K' is chosen such that

$$\min_{K'} \left(\frac{\sum_{j=1}^{K'} \widehat{\lambda}_j}{\sum_{j=1}^{K} \widehat{\lambda}_j} \right) = \min_{K'} \left(\frac{\sum_{j=1}^{K'} \widehat{\lambda}_j}{K} \right) > 1 - \alpha. \tag{A.10}$$

Scott (1992) suggests setting $\alpha = 0.05$. For the data used in this study, with $\boldsymbol{X}$ defined as in (4.1), the first five principal components of $\boldsymbol{X}^*$ contain 95.7, 97.6, and 96.2 percent of the independent nonlinear information in $\boldsymbol{X}^*$ for each year 1985, 1989, and 1994, respectively, and so we chose $K' = 5$.

Partitioning $\boldsymbol{T}$ so that $\boldsymbol{T} = [\,\boldsymbol{T}_1 \quad \boldsymbol{T}_2\,]$, where $\boldsymbol{T}_1$ is $(N \times K')$, we define the $(N \times K')$ matrix

$$\boldsymbol{V} = \boldsymbol{T}_1 (\mathrm{diag}\, \widehat{\boldsymbol{S}}_{K'})^{-1/2}, \tag{A.11}$$

where $\widehat{\boldsymbol{S}}_{K'}$ is the $(K' \times K')$ covariance matrix

$$\widehat{\boldsymbol{S}}_{K'} = N^{-1} \left(\boldsymbol{T}_1 - \overrightarrow{\mathbf{1}}\,\overline{\boldsymbol{T}}_1 \right)' \left(\boldsymbol{T}_1 - \overrightarrow{\mathbf{1}}\,\overline{\boldsymbol{T}}_1 \right). \tag{A.12}$$

Thus the columns of the $\boldsymbol{V}$ each have unit variance and are uncorrelated.

Rather than using $\widehat{m}_h(\boldsymbol{x})$ in (A.6) to estimate the conditional mean function appearing in (4.2), we redefine the model as

$$C_i^* = r(\boldsymbol{V}_i) + \xi_i, \qquad i = 1, \ldots, N, \tag{A.13}$$

where ξ_i is an independent stochastic error term. Analagous to (4.4) and (A.1)–(A.6), the conditional mean function $r(\cdot)$ can be estimated by

$$\widehat{r}_h(\boldsymbol{v}) = \frac{\sum_{i=1}^{N} \underline{\mathcal{K}}_h(\boldsymbol{v} - \boldsymbol{V}_i) C_i^*}{\sum_{i=1}^{N} \underline{\mathcal{K}}_h(\boldsymbol{v} - \boldsymbol{V}_i)}. \tag{A.14}$$

These techniques transform the mapping $\boldsymbol{C} \leftarrow \boldsymbol{X}$ to $\boldsymbol{C}^* \leftarrow \boldsymbol{V}$, reducing the dimensionality of the regression problem to an acceptable level given our sample sizes. Computing $\widehat{C}^* = \widehat{r}_h(\boldsymbol{v})$ using (A.14) involves straightforward numerical computations. Setting $\widehat{C} = \exp \widehat{C}^*$ gives an estimate of cost that can be substituted into any of the measures described in section two to examine returns to scale.

Bandwidth Selection:

The parameter h appearing in the above expressions represents the bandwidth of the kernel estimator, and determines the degree of smoothness of $\widehat{r}_h(\boldsymbol{v})$. At an observation $\boldsymbol{V}_i$, $\lim_{h \to 0} \widehat{r}_h(\boldsymbol{V}_i) = C_i^*$, while at an arbitrary point $\boldsymbol{v}$, $\lim_{h \to \infty} \widehat{r}_h(\boldsymbol{V}_i) = N^{-1} \sum_{i=1}^{N} C_i^*$. Schuster (1972) proves for the univariate case that if h is chosen such that $h = h(N) \to 0$ and $Nh \to \infty$ as $N \to \infty$, then $\widehat{r}_h(v) \xrightarrow{\mathrm{P}} r(v)$, *i.e.*, $\widehat{r}_h(v)$ is a consistent estimator of $r(v)$. It is straightforward to extend this result to the multivariate case.

Härdle and Linton (1994) observe that a bandwidth sequence $\{h_N^*\}$ may be considered asymptotically optimal with respect to some performance criterion $Q(h)$ if

$$\frac{Q(h^*)}{\inf\limits_{h \in H_N} Q(h)} \xrightarrow{\mathrm{P}} 1 \tag{A.15}$$

as $n \to \infty$, where H_N is the range of permissible bandwidths. There are a number of choices for the optimality criterion $Q(h)$. For example, if interest lies in the quadratic loss of the conditional mean estimated at a single point $\boldsymbol{v}$, the appropriate optimality criterion would be the mean square error, denoted $\mathrm{MSE}[\widehat{r}_h(\boldsymbol{v})]$. Alternatively, the integrated mean square error

$$\mathrm{IMSE}(h) = \int \mathrm{MSE}[\widehat{r}_h(\boldsymbol{v})] f(\boldsymbol{v})\, d\boldsymbol{v} \tag{A.16}$$

gives a measure of global performance. Since the goal in the present study is to examine scale efficiencies at observed data points along the estimated cost function, we use the in-sample version of IMSE, *i.e.,* the averaged squared error (ASE)

$$\text{ASE}(h) = N^{-1} \sum_{i=1}^{N} [\widehat{r}_h(\boldsymbol{V}_i) - r_h(\boldsymbol{V}_i)]^2 \tag{A.17}$$

as the optimality criterion $Q(h)$.[22]

Unfortunately, the ASE in (A.17) involves the unknown term $r_h(\boldsymbol{V}_i)$, and therefore must be approximated. A naive approach would be to substitute the observations C_i^* for the unknown values $r_h(\boldsymbol{V}_i)$. But the resulting quantity would then make use of each observation twice, since C_i^* is used in $\widehat{r}_h(\boldsymbol{V}_i)$, and could be made arbitrarily small by taking $h \to 0$. This problem can be avoided by removing the jth observation from the conditional mean estimate, by computing

$$\widehat{r}_{hi}(\boldsymbol{V}_i) = \frac{\sum_{\substack{\ell=1 \\ \ell \neq i}}^{N} \underline{\mathcal{K}}_h(\boldsymbol{V}_i - \boldsymbol{V}_\ell) C_\ell^*}{\sum_{\substack{\ell=1 \\ \ell \neq i}}^{N} \underline{\mathcal{K}}_h(\boldsymbol{V}_i - \boldsymbol{V}_\ell)}, \tag{A.18}$$

and then defining the crossvalidation function

$$CV(h) = N^{-1} \sum_{i=1}^{N} [\widehat{r}_{hi}(\boldsymbol{V}_i) - C_i^*]^2, \tag{A.19}$$

which may be minimized with respect to h to yield an optimal (by this particular criterion) bandwidth.[23]

---

[22] Härdle and Linton (1994) append a weighting funciton $\pi(\boldsymbol{V}_i)$ to the right-hand side of (A.17) to downweight observations in the tails of the distribution of the $\boldsymbol{V}_i$s, thereby reducing boundary effects discussed by Müller (1988). Since the emphasis here is on finding the scale-efficient size of banks, which presumably will not be near the boundary of the $\boldsymbol{V}_i$s, and since as a practical matter it is not clear what an appropriate weighting function would be, we do not weight the observations.

[23] The crossvalidation function $CV(h)$ in (A.19) is well-behaved only in some neighborhood of the optimal bandwidth. Thus, for the K' dimensional case, we take the minimum over $H_N = [0.25\widetilde{h}, 2.0\widetilde{h}]$, where $\widetilde{h} = \left(\frac{4}{2K'+1}\right)^{\frac{1}{K'+1}} N^{-\frac{1}{K'+1}}$. Härdle and Marron (1985) demonstrate that under certain conditions, a bandwidth h chosen to minimize $CV(h)$ is asymptotically optimal with respect to ASE and IMSE.

Bootstrap Estimation of Confidence Intervals:

Bootstrap estimation of confidence interals for our various measures amounts to obtaining bootstrap estimates $\widetilde{C}$, and then approximating the distribution of $\widehat{C} - C$ by $\widetilde{C} - \widehat{C}$. While it would be tempting to resample the rows of $[\boldsymbol{X} \quad \boldsymbol{C}]$, doing so would lead to inconsistent bootstrap estimates. We avoid this problem by using the wild bootstrap, which resamples residuals for observations $i = 1, \ldots, N$ from a two-point distribution uniquely determined so that the first three moments of the resampled residuals will equal 0, $\widehat{\xi}_i^2$, and $\widehat{\xi}_i^3$.[24] This is accomplished by computing residuals

$$\widehat{\xi}_i = C_i^* - \widehat{r}_h(\boldsymbol{v}) \tag{A.20}$$

using (A.13)–(A.14). Then for observation i, the resampled residual $\widetilde{\xi}_i$ equals $\widehat{\xi}_i(1-\sqrt{5})/2$ with probability $\gamma = (5+\sqrt{5})/10$, and equals $\widehat{\xi}_i(1+\sqrt{5})/2$ with probability $(1-\gamma)$. Then new observations

$$\widetilde{C}_i^* = \widehat{r}_g(\boldsymbol{v}) + \widetilde{\xi}_i \tag{A.21}$$

are defined, where the bandwidth g slightly oversmooths the data (*i.e.,* is larger than h; see Härdle and Marron, 1991, for discussion. We set $g = 1.5h$. The choice of g could be refined somewhat, but this problem is beyond the scope of this paper). Then the kernel smoother in (A.14) is applied to the simulated data $[\boldsymbol{V} \quad \widetilde{\boldsymbol{C}}^*]$, yielding the bootstrap estimates $\widetilde{C} = \exp \widetilde{C}^* = \exp \widetilde{r}_h(\boldsymbol{v})$. Repeating this process B times yields a set of estimates $\left\{\widetilde{C}\right\}_{b=1}^{B}$. Substituting these bootstrap values for C into (2.5), 2.7), and (2.9) yields bootstrap values $\left\{\widetilde{\mathcal{S}}_b(\theta|\boldsymbol{y})\right\}_{b=1}^{B}$, $\left\{\widetilde{\mathcal{E}}_b(\theta|\boldsymbol{y}^a, \boldsymbol{y}^b)\right\}_{b=1}^{B}$, and $\left\{\widetilde{\mathcal{A}}_b(\theta|\boldsymbol{y}^a, \boldsymbol{y}^b)\right\}_{b=1}^{B}$. Each of these sets, when suitably centered on the original estimates, then approximates the sampling distributions of the original statistics $\widehat{\mathcal{S}}(\theta|\boldsymbol{y})$, $\widehat{\mathcal{E}}(\theta|\boldsymbol{y}^a, \boldsymbol{y}^b)$, and $\widehat{\mathcal{A}}(\theta|\boldsymbol{y}^a, \boldsymbol{y}^b)$, respectively.

In the case of $\widehat{\mathcal{S}}(\theta|\boldsymbol{y})$, we use the Bonferroni inequality to construct 95 percent simultaneous confidence intervals in Figure 3 at values $\log\theta = -3.5, -2.5, \ldots, 4.5$. Thus,

[24] For a discussion of the inconsistency of the naive bootstrap and the alternative wild bootstrap, see Härdle (1990), Härdle and Marron (1991), Mammen (1991), and Härdle and Mammen (1993). Cao-Abad (1991) discusses convergence rates for the wild bootstrap.

for each value of $\log\theta$, we sort the values in $\left\{\widetilde{S}_b(\theta|\boldsymbol{y})\right\}_{b=1}^{B}$ by algebraic value, and then take the $((0.05/2)/8\times 100)$th and the $((1-(0.05/2)/8)\times 100)$th quantiles as the 95 percent simultaneous confidence limits. Confidence intervals are obtained similarly for Figures 4 and 5, but since we only want pointwise intervals, we do not divide by the factor 8 in the previous expressions. For the confidence intervals in Figure 3, we set $B = 5000$, since we are in effect estimating the extreme tails of the sampling distributions; in Figures 4 and 5, we set $B = 1000$ (see Hall, 1986).

APPENDIX B

We first estimate the translog cost function in (4.3) separately for each year, using observations that contain no zero values for any of the output variables.[25] While it is straightforward to estimate (4.3) using ordinary least squares, statistical efficiency can be improved by incorporating information from factor share equations

$$\begin{aligned} S_{ik} =& \beta_k + \sum_{j=1}^{3} \delta_{jk}\log(p_{ij}/p_{i4}) + \delta_{kk}\log(p_{ik}/p_{i4}) + \\ & \sum_{j=1}^{4} \tau_{jk}\log y_{ij} + \psi_k \log x_{i5} + \xi_{ik}, \qquad k = 1,2,3, \end{aligned} \tag{B.1}$$

where for estimation purposes S_{ik} is computed as $p_{ik}x_{ik}/C_i$. The right-hand side of (B.1) is obtained from the derivative $\partial \log C_i/\partial \log(p_{ik}/p_{i4})$. The equations represented by (B.1), together with the translog cost function in (4.3), comprise a system of seemingly unrelated regressions. We estimate the parameters using Feasible generalized least squares (FGLS) while imposing the cross-equation parameter restrictions.[26]

To test the translog specification of bank costs, we divided the subsamples for each year into four further subsamples according to asset-size:

[25]The numbers of observations with no zero values for any of the output variables are 12,622, 11,363, and 8,924 for 1985, 1989, and 1994, respectively. Alternatively, one could change the zero values to small postive values, which would allow logarithms necessary for the translog specification to be computed. This approach, however, is entirely arbitrary, and the results would likely be sensitive to choice of the small positive value chosen.

[26]The actual parameter estimates are omitted to conserve space, but are available on request.

Category	Size	Observations (1985)	(1989)	(1994)
#1	total assets $\leq$ \$100 million	9,712	8,559	6,551
#2	\$100 million $<$ total assets $\leq$ \$300 million	2,176	2,071	1,731
#3	\$300 million $<$ total assets $\leq$ \$1 billion	577	550	458
#4	total assets $>$ \$1 billion	157	183	184.

If the translog cost function were a correct specification of the cost function, then estimation using each of the four subsamples listed above for a particular year should yield similar parameter estimates. However, reestimating the translog cost function with the factor share equations using data in each of the four size categories listed above yielded some rather large changes in the coefficient estimates within each year.

To test whether the differences in parameter estimates across the various subsamples are jointly significant, we assume normality in the error terms and construct a modified Chow-test statistic:

$$\frac{\left(\mathrm{ESS}_0 - \mathrm{ESS}_1 - \mathrm{ESS}_2 - \mathrm{ESS}_3 - \mathrm{ESS}_4\right)/3K}{\left(\mathrm{ESS}_1 + \mathrm{ESS}_2 + \mathrm{ESS}_3 + \mathrm{ESS}_4\right)/(N-4K)} \sim F_{3K,N-4K} \tag{B.2}$$

where N represents the total number of observations in a particular year, K represents the number of parameters in the translog cost function (45), ESS_0 represents the error sum-of-squares for the complete regression, and ESS_j, $j = 1, 2, 3, 4$ denotes the error sum-of-squares for the jth subsample described above. Computing the test statistic for 1985, 1989, and 1994, we obtain values 10.06, 10.50, and 7.90, respectively, allowing us to reject the null hypothesis of no differences in the parameter vectors at well over 99.9 percent significance for each year.

To check whether this result might be attributable to only one or two asset-size categories, we tested all pairwise differences in parameter vectors estimated from the four subsamples using the Wald statistic

$$W_{jk} = \left(\widehat{\boldsymbol{Z}}_j - \widehat{\boldsymbol{Z}}_k\right)' \left(\widehat{\boldsymbol{S}}_j + \widehat{\boldsymbol{S}}_k\right)^{-1} \left(\widehat{\boldsymbol{Z}}_j - \widehat{\boldsymbol{Z}}_k\right) \overset{\mathrm{a}}{\sim} \chi^2(K), \tag{B.3}$$

where $\widehat{\boldsymbol{Z}}_j, \widehat{\boldsymbol{Z}}_k$ denote the vectors of parameter estimates from the jth and kth subsamples, and $\widehat{\boldsymbol{S}}_j, \widehat{\boldsymbol{S}}_k$ denote the FGLS estimates of the covariance matrices for the parameter vectors from the jth and kth subsamples. With the assumption of normality of the error terms, the Wald statistic is asymptotically chi-square distributed with 45 degrees of freedom. For each year, we are able to reject the null hypothesis of no difference in parameter vectors for all pairwise combinations, at any reasonable level of significance (estimated values of the statistic ranged from 244.34 to 2623.30).[27]

[27]The assumption of normally distributed errors in (4.3) may not be justified if banks are cost-inefficient. However, in each instance our tests based on (B.2) and (b.3) lead to rejection of the null hypotheses at far greater than 99.9 percent significance. While it would be straigthforward to bootstrap the distributions of our test statistics to allow for violation of our normality assumption, experience suggests that this would not make a qualitative difference in our conclusions (*e.g.,* see Stahl and Wilson, 1994, 1995, and Haruvy *et al.,* 1997).

REFERENCES

Akhavein, Jalal D., Berger, Allen N., and David B. Humphrey, "The Effects of Megamergers on Efficiency and Prices: Evidence from a Bank Profit Function," forthcoming, *Review of Industrial Organization*, 1996.

Beran, R., and G. Ducharme (1991), *Asymptotic Theory for Bootstrap Methods in Statistics,* Montreal: Centre de Reserches Mathematiques, University of Montreal.

Berger, A.N., and D.B. Humphrey (1991), The dominance of inefficiencies over scale and product mix economies in banking, *Journal of Monetary Economics* 28, 117–148.

Berger, Allen N. and Humphrey, David B., "Megamergers in Banking and the Use of Cost Efficiency as an Antitrust Defense," *The Antitrust Bulletin* 37 (Fall 1992), pp. 541-600.

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

Boyd, John H. and Graham, Stanley L., "Investigating the Banking Consolidation Trend," *Quarterly Review*, Federal Reserve Bank of Minneapolis (Spring 1991), pp. 3-15.

Cao-Abad, R. (1991), Rate of convergence for the wild bootstrap in nonparametric regression, *Annals of Statistics* 19, 2226–2231.

Clark, Jeffrey A., "Economic Cost, Scale Efficiency, and Competitive Viability in Banking," *Journal of Money, Credit, and Banking* 28 (August 1996), pp. 342-64.

Ferrier, G.D. and C.A.K. Lovell (1990), Measuring cost efficiency in banking: Econometric and linear programming evidence, *Journal of Econometrics* 46, 229-245.

Gasser, T. and H.G. Müller (1979), "Kernel Estimation of Regression Functions," in *Smooting Techniques for Curve Estimation,* ed. by T. Gasser and M. Rosenblatt, Berlin: Springer-Verlag, pp. 23–68.

Gasser, T. and H.G. Müller (1984), Estimating regression functions and their derivatives by the kernel method, *Scandinavian Journal of Statistics* 11, 171–185.

Gropper, Daniel M., "An Empirical Investigation of Changes in Scale Economies for the Commercial Banking Firm, 1979-1986," *Journal of Money, Credit, and Banking* 23 (November 1991), pp. 718-27.

Hall, P. (1986), On the number of bootstrap simulations required to construct a confidence interval, *The Annals of Statistics* 14, 1453–1462.

Härdle, W. (1990), *Applied Nonparametric Regression,* Cambridge: Cambridge University Press.

Härdle, W. and R.J. Carroll (1990), Biased crossvalidation for a kernel regression estimator and its derivatives, *Österreichische Zeitschrift für Statistik und Informatik* 20, 53–64.

Härdle, W. and O. Linton (1994), "Applied Nonparametric Methods," discussion paper

#9402, Institut de Statistique, Université Catholique de Louvain, Louvain-la-Neuve, Belgium.

Härdle, W. and E. Mammen (1993), Comparing nonparametric versus parametric regression fits, *Annals of Statistics* 21, 1926–1947.

Härdle, W. and J.S. Marron (1985), Optimal bandwidth selection in nonparametric regression function estimation, *Annals of Statistics* 13, 1465–1481.

Härdle, W. and J.S. Marron (1991), Bootstrap simulataneous error bars for nonparametric regression, *Annals of Statistics* 19, 778–796.

Haruvy, E., D.O. Stahl, and P.W. Wilson (1997), "Modelling and Testing for Heterogeneity in Observed Strategic Behavior," unpublished working paper, Department of Economics, University of Texas at Austin, USA.

Humphrey, David B., "Why Do Estimates of Bank Scale Economies Differ?" *Economic Review*, Federal Reserve Bank of Richmond, (September/October 1990), pp. 38-50.

Hunter, William C. and Timme, Stephen G., "Core Deposits and Physical Capital: A Reexamination of Bank Scale Economies and Efficiency with Quasi-Fixed Inputs," *Journal of Money, Credit, and Banking* 27 (February 1995), pp. 165-85.

Hunter, William C., Timme, Stephen G., and Won Keun Yang, "An Examination of Cost Subadditivity and Multiproduct Production in Large U.S. Commercial Banks," *Journal of Money, Credit, and Banking* 22 (November 1990), pp. 504-25.

Jagtiani, Julapa and Khanthavit, Anya, "Scale and Scope Economies at Large Banks: Including Off-Balance Sheet Products and Regulatory Effects (1984-1991)," *Journal of Banking and Finance* 20 (1996), pp. 1271-87.

Kaparakis, E.I., S.M. Miller, and A. Noulas (1994), Short-run cost inefficiency of commercial banks: A flexible stochastic frontier approach, *Journal of Money, Credit and Banking* 26, 875-893.

Mammen, E. (1991), *When Does Bootstrap Work? Asymptotic Results and Simulations,* Berlin: Springer-Verlag.

McAllister, P.H., and D. McManus (1993), Resolving the scale efficiency puzzle in banking, *Journal of Banking and Finance* 17, 389–405.

Mitchell, Karlyn and Onvural, Nur M., "Economies of Scale and Scope at Large Commercial Banks: Evidence from the Fourier Flexible Functional Form," *Journal of Money, Credit, and Banking* 28 (May 1996), pp. 178-99.

Müller, H.G. (1988), *Nonparametric Regression Analysis of Longitudinal Data,* Berlin: Springer-Verlag.

Nadarya, E.A. (1964), On estimating regression, *Theory of Probability and its Applications* 10, 186–190.

Schuster, E.F. (1972), Joint asymptotic distribution of the estimated regression function at a finite number of distinct points, *Annals of Mathematical Statistics* 43, 84–88.

Scott, D.W. (1992), *Multivariate Density Estimation: Theory, Practice, and Visualization,* New York: John Wiley and Sons, Inc.

Silverman, B.W. (1986), *Density Estimation for Statistics and Data Analysis,* London: Chapman and Hall, Ltd.

Stahl, D.O., and P.W. Wilson (1994), Experimental evidence on players' models of other players, *Journal of Economic Behavior and Organization* 25, 309–327.

Stahl, D.O., and P.W. Wilson (1995), On players' models of other players: Theory and experimental evidence, *Games and Economic Behavior* 10, 218–254.

Watson, G. (1964), Smooth regression analysis, *Sankhya Series A* 26, 359–372.

Wheelock, D.C., and P.W. Wilson (1996), "Technical Progress, Inefficiency, and Productivity Change in US Banking, 1984–1993," unpublished working paper, Department of Economics, University of Texas.

Figure 1
Graphical Representation of Comptetive Viability Measures

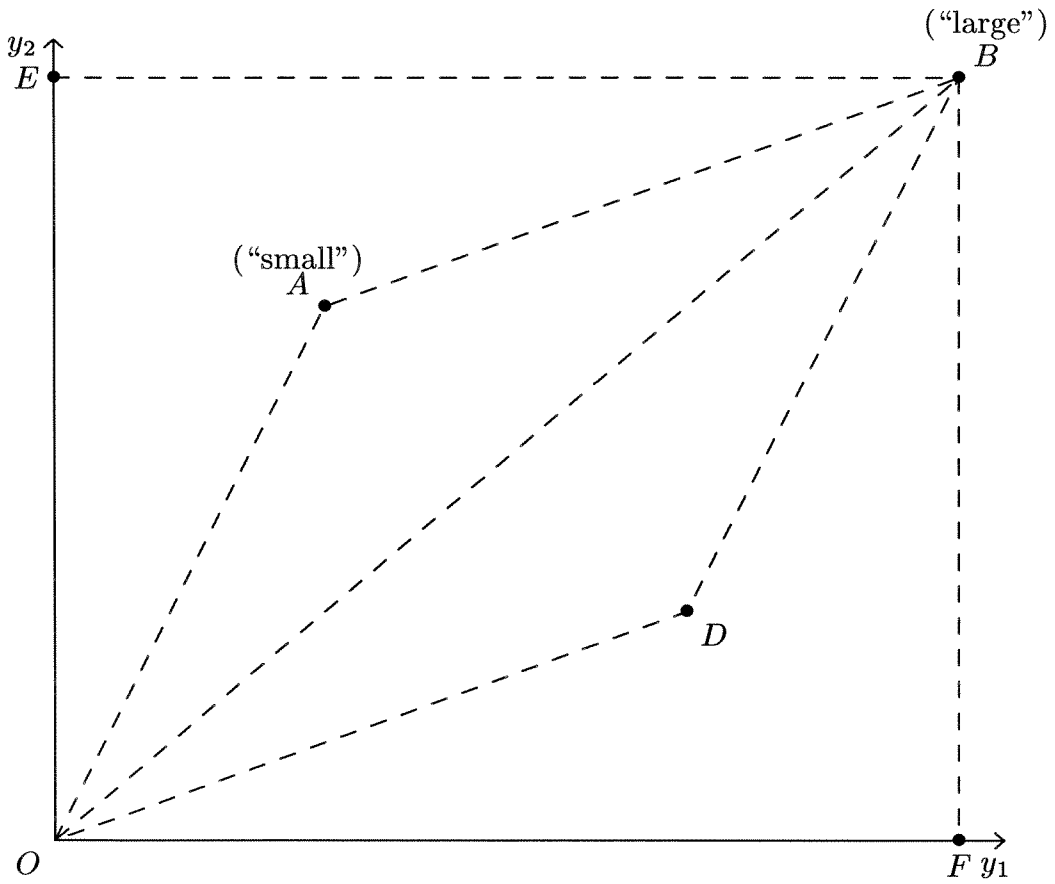

Figure 2
Density of log(*ASSETS*)

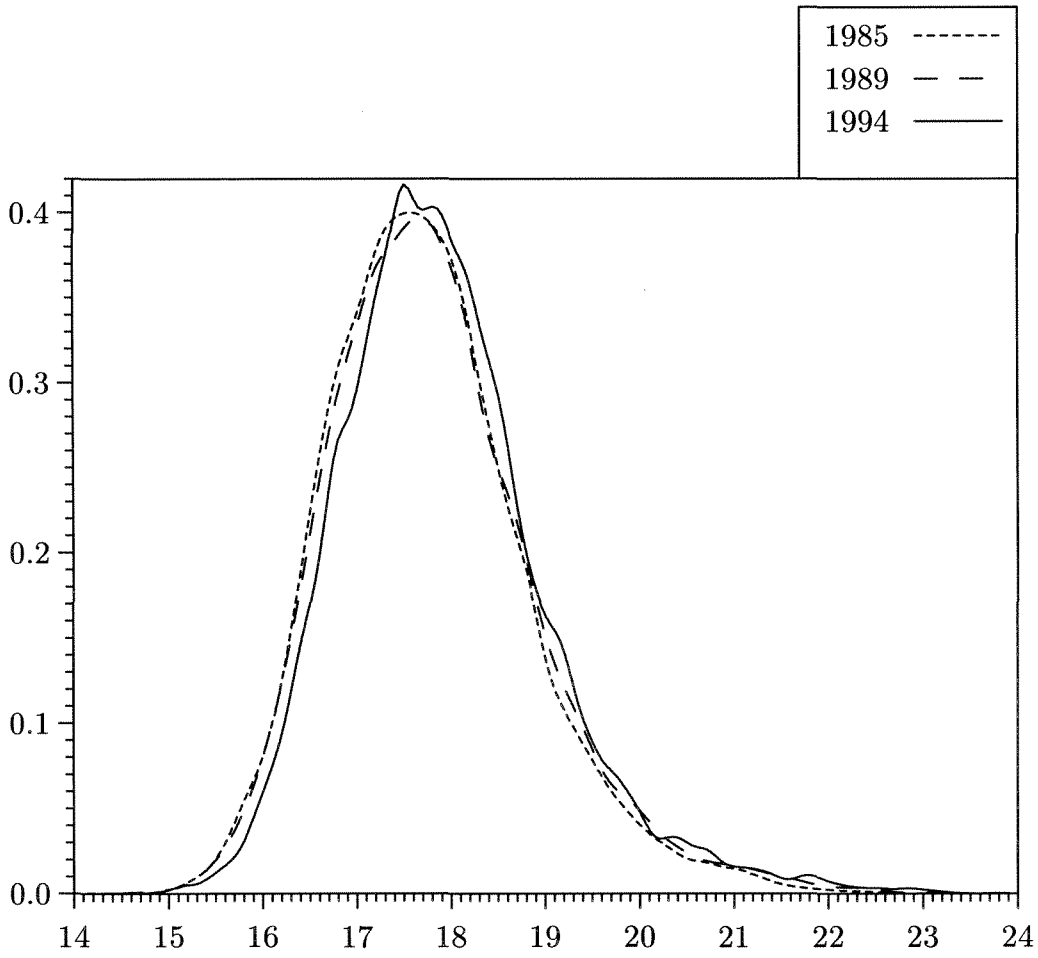

Figure 3
Ray Scale Economies

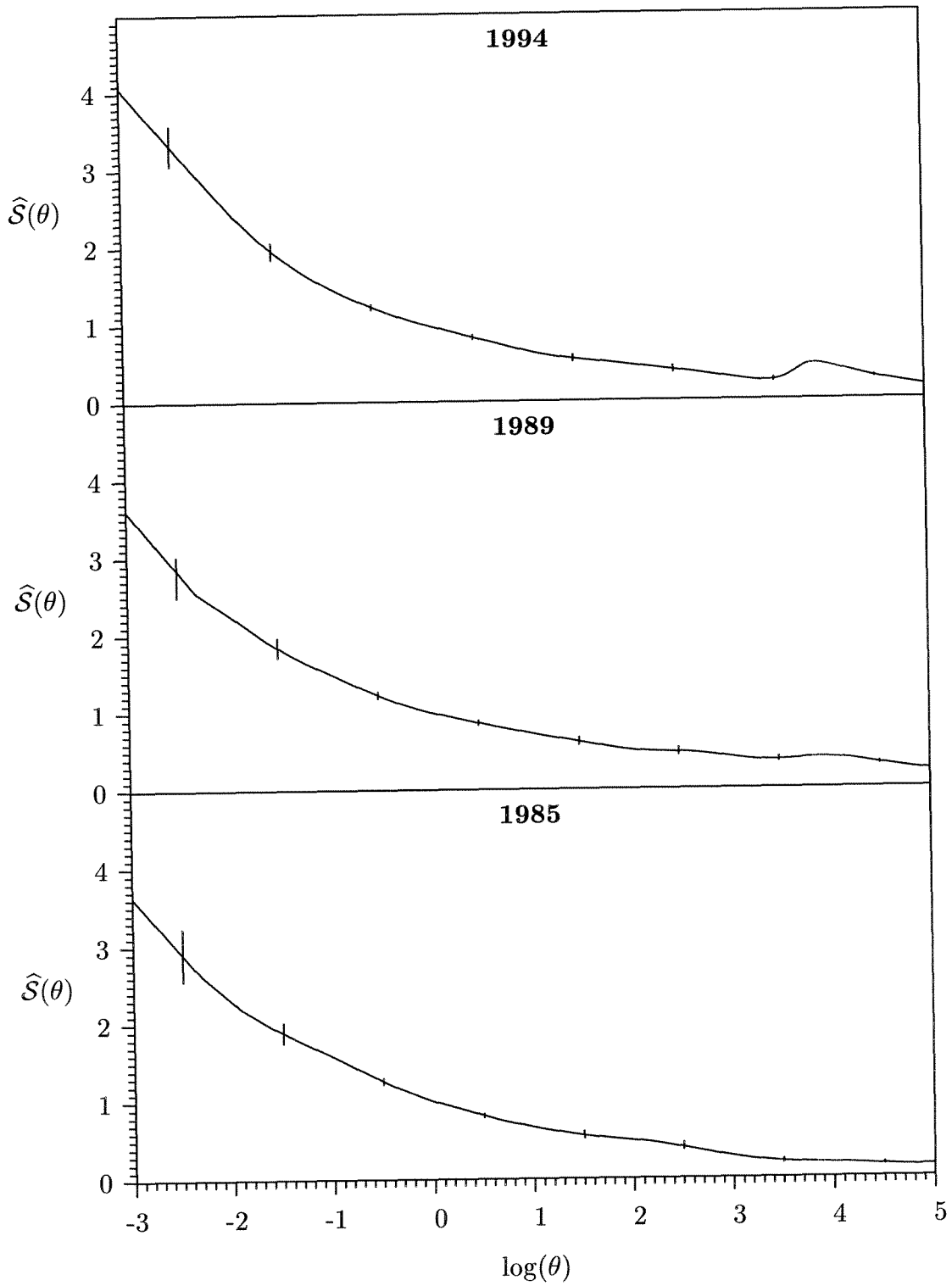

Figure 4
Expansion Path Scale Economies

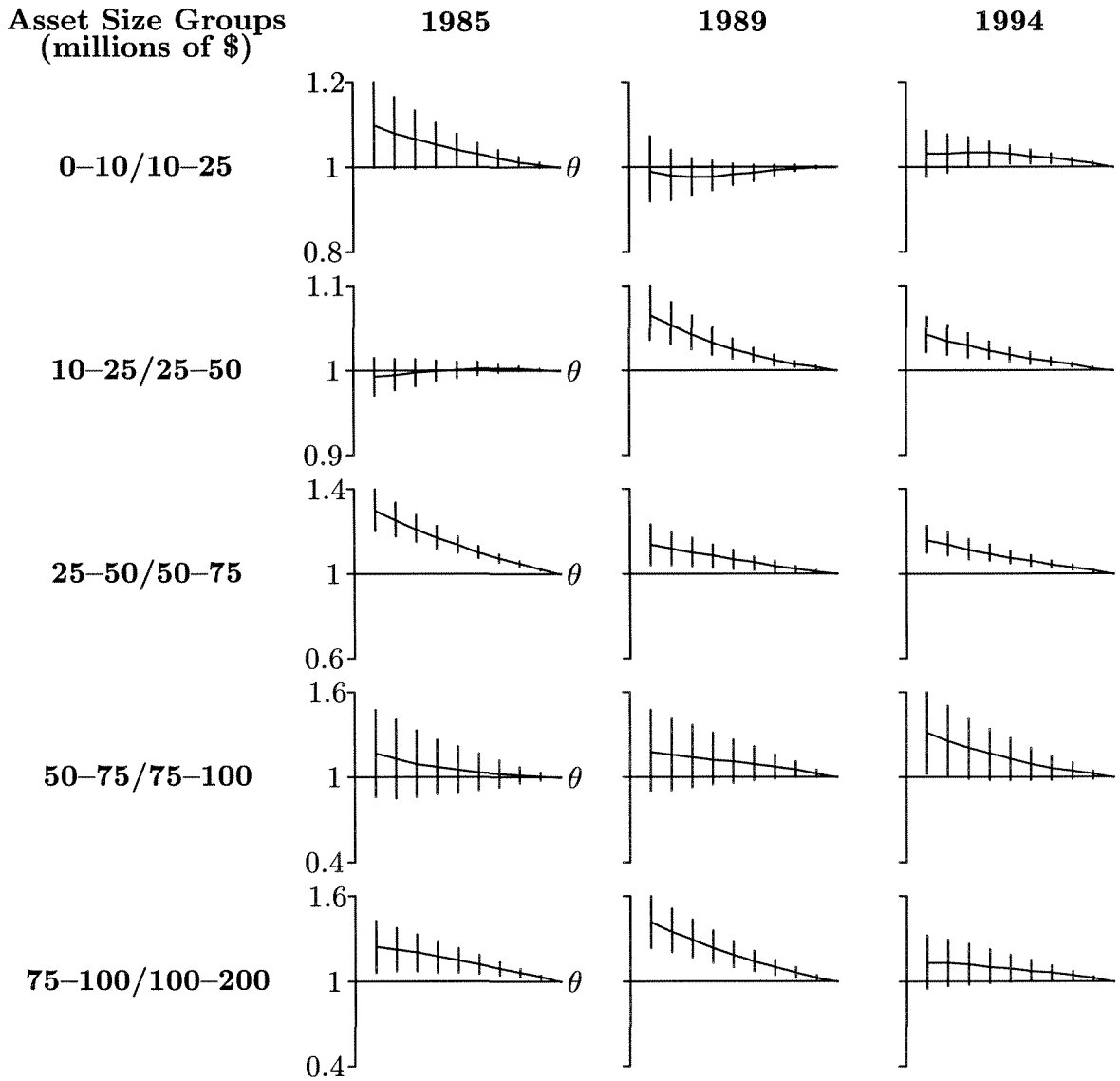

Figure 4 (continued)

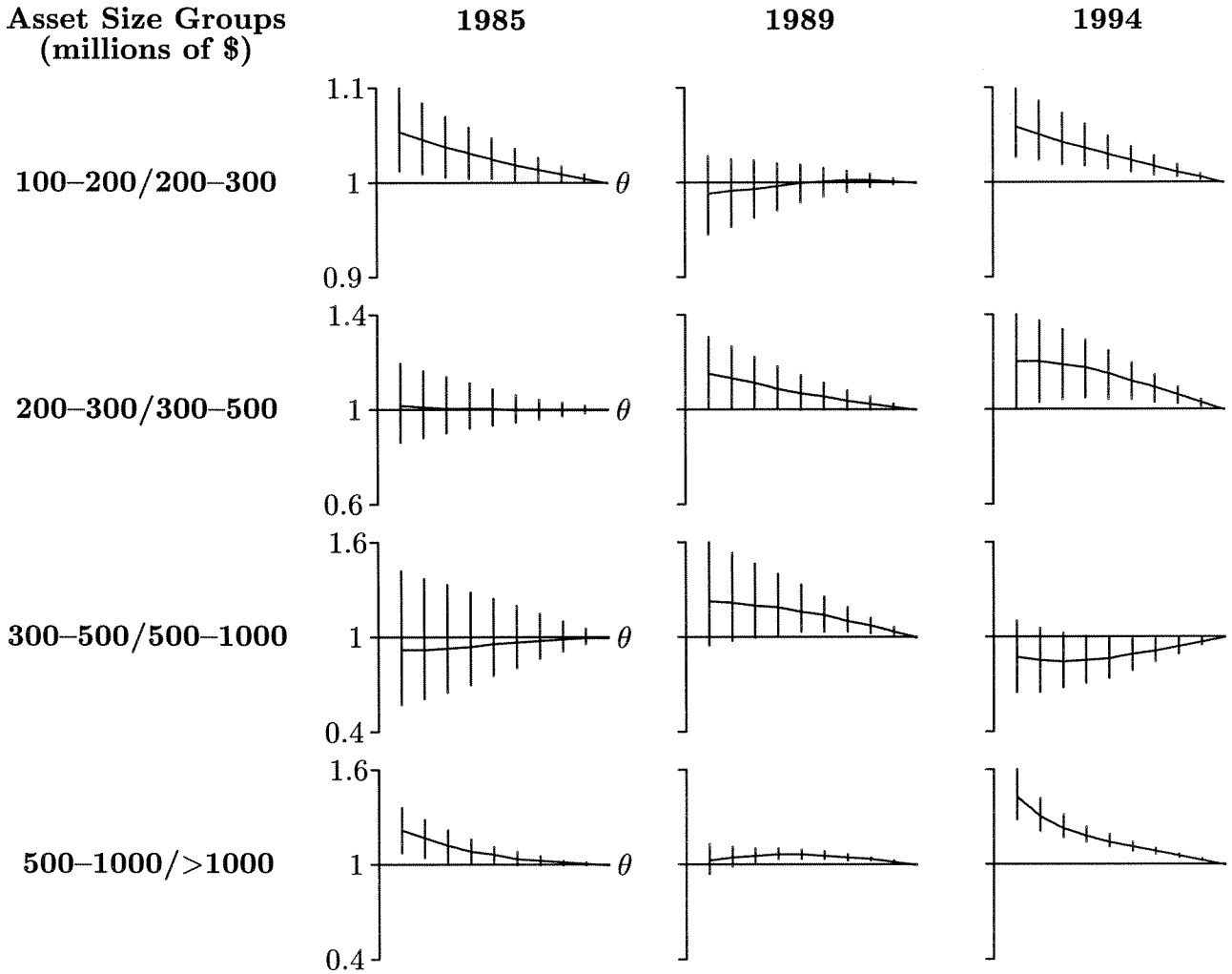

Figure 5
Expansion Path Subadditivity

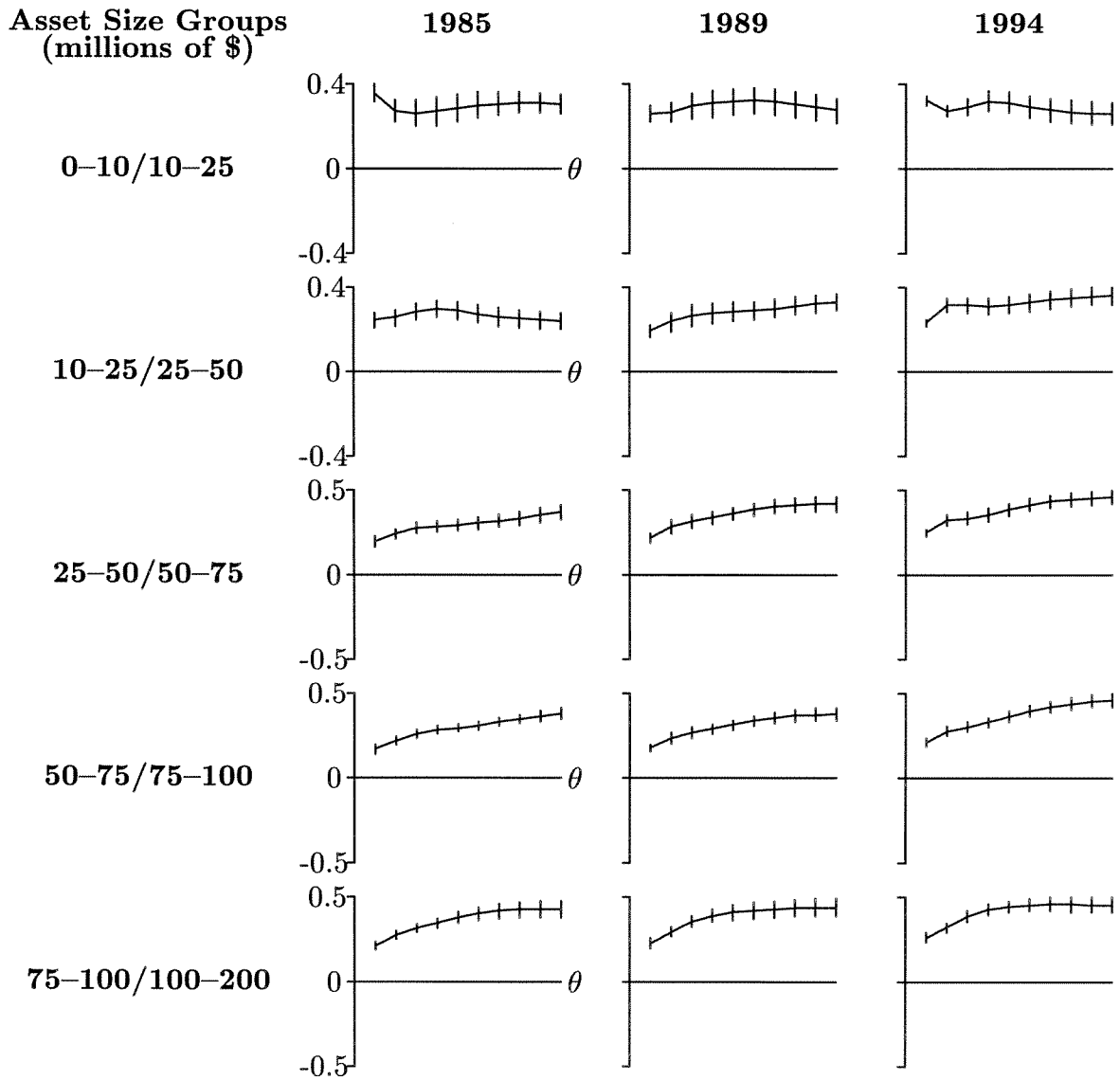

Figure 5 (continued)