FEDERAL RESERVE BANK of ST. LOUIS

RESEARCH DIVISION

Working Paper Series

Do Large Banks Have Lower Costs? New Estimates of Returns to Scale for U.S. Banks

David C. Wheelock
and
Paul W. Wilson

Working Paper 2009-054E
https://doi.org/10.20955/wp.2009.054

May 2011

FEDERAL RESERVE BANK OF ST. LOUIS
Research Division
P.O. Box 442
St. Louis, MO 63166

Do Large Banks have Lower Costs? New Estimates of Returns to Scale for U.S. Banks

David C. Wheelock Paul W. Wilson*

May 2011

Abstract

The number of commercial banks in the United States has fallen by more than 50 percent since 1984. This consolidation of the U.S. banking industry and the accompanying large increase in average (and median) bank size have prompted concerns about the effects of consolidation and increasing bank size on market competition and on the number of banks that regulators deem "too–big–to–fail." Agency problems and perverse incentives created by government policies are often cited as reasons why many banks have pursued acquisitions and growth, though bankers often point to economies of scale. This paper presents new estimates of ray-scale and expansion-path scale economies for U.S. banks based on non-parametric local-linear estimation of a model of bank costs. Unlike prior studies that use models with restrictive parametric assumptions or limited samples, our methodology is fully non-parametric and we estimate returns to scale for all U.S. banks over the period 1984–2006. Our estimates indicate that as recently as 2006, most U.S. banks faced *increasing* returns to scale, suggesting that scale economies are a plausible (but not necessarily only) reason for the growth in average bank size and that the tendency toward increasing scale is likely to continue unless checked by government intervention.

*Wheelock: Research Department, Federal Reserve Bank of St. Louis, P.O. Box 442, St. Louis, MO 63166–0442; wheelock@stls.frb.org. Wilson: The John E. Walker Department of Economics, 222 Sirrine Hall, Clemson University, Clemson, South Carolina 29634–1309, USA; email pww@clemson.edu. This research was conducted while Wilson was a visiting scholar in the Research Department of the Federal Reserve Bank of St. Louis. We thank the Cyber Infrastructure Technology Integration group at Clemson University for operating the Palmetto cluster used for our computations; we are especially grateful to Barr von Oehsen for technical support and advice. We thank Craig Aubuchon, Heidi Beyer and David Lopez for research assistance, and we thank the editor, Bob DeYoung, and an anonymous referee for comments on a previous version of this paper. The views expressed in this paper do not necessarily reflect official positions of the Federal Reserve Bank of St. Louis or the Federal Reserve System. *JEL* classification nos.: G21, L11, C12, C13, C14. Keywords: banks, returns to scale, scale economies, non-parametric, regression.

1 Introduction

The past 25 years have witnessed a consolidation of the U.S. banking industry on a scale not seen since the Great Depression. Between 1984 and 2008, the number of U.S. commercial banks fell by more than 50 percent, from 14,482 to 7,086. Over the same period, the average size of U.S. banks increased five-fold in terms of inflation-adjusted total assets. Bank executives and industry analysts contend that changes in regulation and advances in information-processing technology have encouraged banks to grow larger, and often point to economies of scale to justify bank mergers.[1] Critics contend that consolidation has had a deleterious impact on market competition, however, and that the increase in the size of banks reflects agency problems and government policies that disproportionately benefit large banks. In particular, the recent financial crisis has brought forward concerns about banks that regulators deem "too–big–to–fail" in the sense that their failure would pose serious systemic risks, which has prompted calls for regulatory limits on bank size (e.g., Reich, 2008; O'Driscoll, 2009). However, Stern and Feldman (2009) contend that policymakers should consider the loss of any scale benefits when determining the net benefit of limiting the size of banks. Hence, assessment of the extent of scale economies is important for gauging the costs and benefits of any policy intervention to control the size of banks.

Despite the claims of bankers, with few exceptions, researchers have found little evidence of significant scale economies in banking. Early studies found that banks exhaust scale economies at \$100-\$200 million of total assets, suggesting little cost savings are generated through either bank mergers or internally-generated growth. However, much of the early research on scale economies in banking involved the estimation of parametric cost functions that fail basic specification tests or models that fail to capture key features of bank production.[2] Differences in estimates of scale economies between earlier and more recent studies may also partly reflect the removal of branching restrictions and other changes in regulation that have made it less costly for banks to become large in recent years (Mester, 2005). Further, over time, technological advances may have favored larger banks and thereby affected returns to scale. Information processing equipment and software entail rather high fixed

[1] For example, see Bach (2001), Thompson (2002), and Rieker (2006). Berger (2003) describes the myriad advances in information and financial technology and changes in regulation affecting the banking industry over these years, and discusses their implications for banks of different sizes.

[2] McAllister and McManus (1993) and Mester (2005) survey the older literature on scale economies.

costs. Moreover, reductions in the cost of acquiring quantifiable information about potential borrowers have eroded some of the benefits of small scale and close proximity to borrowers that enabled small banks traditionally to out-compete larger banks for some customers, such as small businesses (see Petersen and Rajan, 2002; Berger, 2003; and Bernanke, 2006).

Recent research has found considerably more evidence of scale economies in banking. For example, McAllister and McManus (1993) and Wheelock and Wilson (2001) find that banks face increasing returns to scale up to at least $500 million of total assets. Both studies use non-parametric and semi-non-parametric methods that avoid the problem of specifying *a priori* a particular functional cost relationship to be estimated. Likewise, studies that incorporate banks' risk preferences and financial capital into models of bank production find more evidence of increasing returns to scale than studies that ignore these effects. For example, Hughes et al. (2001) estimate returns to scale within the context of a value maximization model that explicitly incorporates the capital structure and risk-taking preferences of individual banks. Based on a sample of 441 top-tier bank holding companies in 1994, Hughes et al. (2001) find that large banks face significant scale economies that increase with bank size.

Feng and Serilitis (2009) also find that large banks operate under increasing returns to scale. That study derives estimates of returns to scale from Bayesian estimation of a translog output distance function, rather than from a cost function. This approach has the advantage of avoiding the use of input prices, which may be subject to considerable measurement error. Based on a sample of 292 banks with at least $1 billion of total assets during 2000-05, Feng and Serilitis (2009) find that all banks exhibit increasing returns to scale. As the study acknowledges, however, the translog specification is suitable only for samples composed of relatively homogeneous firms. A different approach is required for estimating scale economies across banks of widely differing sizes.

The present paper reports new estimates of returns to scale for banks throughout the distribution of observed bank sizes for 1984–2006. We estimate returns to scale in a cost framework, which provides evidence on whether society's resources are allocated efficiently by addressing directly the long-controversial question whether banks can lower their average costs by increasing scale. Although bankers often claim that banks can lower costs by expanding in size, many policymakers and academics remain skeptical (see, e.g., Haldane,

2010 and Greenspan, 2010).[3]

We use a non-parametric local-linear estimator to estimate the cost relationship for commercial banks and to derive estimates of ray-scale and expansion-path scale economies, and thereby avoid the potential for specification error associated with parametric estimation.[4] Non-parametric estimators are subject to the well-known "curse of dimensionality," and large sample sizes are required because of the slow convergence rates of non-parametric estimators. We employ principal components techniques to reduce the dimensions of our empirical model. Further, we estimate the model using two large datasets. Sample #1 consists of 887,369 quarterly observations on all U.S. commercial banks for 1984-2006. Sample #2 consists of 868,647 quarterly observations on all commercial banks for 1984-2000, and on a mix of commercial banks and bank holding companies comprising the largest top-tier banking organizations for 2001-06.[5] In consisting of data for only commercial banks, the first sample has the advantages of being both somewhat larger and more homogeneous than the second dataset. Further, because most bank holding companies had consolidated their banking assets by 2006 so that the lead bank comprised all or nearly all of the banking assets of the entire holding company, the first sample includes some banks that were as large as nearly all bank holding companies. Only two bank holding companies were larger than the largest bank in our first sample in the fourth quarter of 2006, and only four holding companies were larger than that bank as of the first quarter of 2010. Nonetheless, by including the very largest banking organizations, the second sample is useful for estimating returns to scale

[3] Although our approach can provide information about the extent of cost economies in banking, it does not address such questions as whether increased bank scale affects the costs incurred by bank customers, the risks incurred by banks, or the risks that the banking system may impose on society more broadly. Furthermore, Hughes et al. (2001) and other studies report evidence suggesting that banks do not, in fact, minimize cost. This evidence is consistent with their pursuit of other objectives, such as value maximization, as well as with possible agency problems between bank managers and shareholders, or simply inefficiency. Following the approach of Hughes et al. (2001), we also find evidence that many banks do not use optimal amounts of capital input (results given in a separate Appendix D, available from the authors upon request).

[4] Like McAllister and McManus (1993) and Wheelock and Wilson (2001), we tested and easily rejected the stability of the parameters of a translog cost function estimated across banks of different sizes (a description of the test and results are given in Section 3 and a separate Appendix A, which is available from the authors upon request).

[5] Specifically, for 2001-06, sample #2 is comprised of observations on all independent commercial banks, commercial banks owned by non-reporting holding companies, and consolidated data for reporting top-tier holding companies. Due to changes in reporting requirements for holding companies, data on holding companies for earlier years do not allow construction of some of the variables in our model for periods prior to 2001. In addition, since 2005, only holding companies with at least $500 million of assets have been required to file consolidated financial reports.

throughout the entire range spanned by the data.

For both samples, we incorporate discrete covariates into the estimation procedure to control for whether a commercial bank was owned by a multi-bank holding company, as well as for time and whether a bank was located in a state that restricted branch banking. Multi-bank holding companies often provide services, such as advertising and access to ATM networks, to their subsidiaries that may cause their cost structures to differ from those of other banks. For the early years of our sample, we also distinguish between banks that operated in states that restricted branching versus those in states with more liberal regulations. Interstate branching was not permitted until 1997, and in prior years several states restricted branching within their state borders. Although all banks have operated under full interstate branching since 1999, for earlier years we produce separate estimates for banks located in states that prohibited branching, permitted some branching, or that permitted state-wide branching. Finally, we use cross-validation techniques to optimize bandwidths and a non-parametric bootstrap procedure for inference. Our large sample size and use of non-parametric estimators results in a substantial computational burden, which we handle using parallel programming techniques and a massively parallel computer (details are given below in Section 4).

Our estimates reveal that most banking organizations, including the very largest holding companies, operated under increasing returns to scale throughout our sample period. Hence, our findings are consistent with other recent studies that find evidence of significant scale economies for large banking organizations, as well as with the view that industry consolidation has been driven, at least in part, by scale economies. Further, our results have implications for policies intended to limit the size of banks to ensure competitive markets, to reduce the number of banks deemed "too-big-to-fail," or for other purposes. Although there may be benefits to imposing limits on the size of banks, our research points out potential resource costs of such intervention.[6]

[6] Federal banking law restricts individual banking organizations from engaging in most acquisitions that would result in the organization holding more than 10-percent of the nation's total bank deposits, or more than 30-percent of an individual state's total deposits. In addition, the Dodd-Frank Wall Street Reform and Consumer Protection Act prohibits most acquisitions or mergers that would result in the concentration of more than 10 percent of the aggregated liabilities of all financial companies in a single firm. U.S. bank regulators and the Department of Justice review all proposed bank mergers and acquisitions and will deny transactions that would result in excessively concentrated banking markets (see Gilbert and Zaretsky, 2003).

The next section presents a model of bank costs and describes the ray-scale and expansion-path measures of scale economies. Section 3 discusses our non-parametric estimator and methods for inference. Section 4 presents estimation results, and Section 5 offers our conclusions.

2 A Model of Bank Costs

2.1 Specifying the Cost Mapping

To estimate scale economies we must first specify a model of bank costs. Two issues are involved: (i) the choice of appropriate variables, and (ii) given those variables, the mapping of output quantities, input prices, and other arguments of the cost relationship.

With regard to variable specification, we define five inputs and five outputs that, with one exception (the measure of off-balance sheet output), are those used by Berger and Mester (2003). Specifically, we define the following output quantities: consumer loans (Y1), business loans (Y2), real estate loans (Y3), securities (Y4), and off-balance sheet items (OFF) consisting of net non-interest income minus service charges on deposits.[7] We define three variable input quantities: (i) purchased funds, consisting of the sum of total time deposits of $100,000 or more, foreign deposits, federal funds purchased, demand notes, trading liabilities, other borrowed money, mortgage indebtedness and obligations under capitalized leases, and subordinated notes and debentures;[8] (ii) core deposits, consisting of total deposits less time deposits of $100,000 or more; and (iii) labor services, measured by the number of full-time equivalent employees on payroll at the end of each quarter. We measure the prices of purchased funds (W1), core deposits (W2), and labor services (W3) by dividing total expenditure on the given input by its quantity. Variable cost (COST) is the sum of expenditures on these three inputs. Finally, we define two fixed netput quantities: physical capital, consisting of premises and other fixed assets (Z1), and financial equity capital (Z2).[9]

[7] Of the commonly used measures of off-balance sheet output, net non-interest income is the most consistently measurable across banks and over time. However, as a net, rather than gross measure of income, it is potentially a biased measure of off-balance sheet output because losses would appear to reduce off-balance sheet output. Data are not reported that would permit calculation of a gross measure of non-interest income. See Clark and Siems, 2002 for discussion of alternative measures of off-balance sheet activity.

[8] This variable is intended to capture non-core deposit (and non-equity) sources of investment funds for the bank. See Berger and Mester (2003) for more detail.

[9] Our treatment of physical and equity capital as quasi-fixed reflects the absence of reliable price data

With the exception of labor input (which is measured as full-time equivalent employees) and off-balance sheet output (which is measured in terms of net flow of income), our inputs and outputs are stocks measured by dollar amounts reported on bank balance sheets, consistent with the widely used "intermediation" model of Sealey and Lindley (1977).

In addition to the variables defined above, we index quarters 1984.Q1 through 2006.Q4 by setting TIME = 1 for 1984.Q1, TIME = 2 for 1984.Q2, ..., TIME = 92 for 2006.Q4. Although TIME is an ordered, categorical variable, we treat it as continuous since it can assume a wide range of possible values. The regulatory environment and the production technology of banking changed a great deal over the 23 years covered by our data; consequently, it seems important to include TIME as an explanatory covariate in the cost function. Two features of our estimation strategy allow a great deal of flexibility. First, because we use a fully non-parametric estimation method, we impose no constraints on how TIME might interact with other explanatory variables. Second, the local nature of our estimator means that when we estimate cost at a particular point in time, observations from distant time periods will have little or no effect on the estimate. By contrast, typical approaches that involve estimation of a fully parametric translog cost function model by OLS or some other estimation procedure are not local in the sense that when cost is estimated at some point in the data space, all observations contribute to the estimate with equal weight. Moreover, the typical approach requires the imposition of a specific functional form a priori for any interactions among explanatory variables.[10]

To control for differences in costs associated with holding company affiliation, we define MBHC = 1 for commercial banks that are owned by a multi-bank holding company (MBHC = 0 otherwise). We also employ three binary variables to control for the different branch banking regimes that existed during our sample period. Full interstate branching has been in effect since 1999. Before then, state laws specified the extent to which banks were permitted to operate branches within individual states. We define (i) STATEWIDE = 1 for banks operating in states that permitted state-wide branching (STATEWIDE = 0 otherwise); (ii) LIMITED = 1 for banks operating in states that permitted limited branching

and is consistent with other recent studies (e.g., Berger and Mester, 2003). Some studies (e.g., Hughes et al., 2001) also include nonperforming loans as an additional source of cost.

[10] The local nature of our estimator is discussed in more detail below in Section 3 and in a separate Appendix B, which is available from the authors upon request.

(LIMITED = 0 otherwise); and (iii) UNIT = 1 for banks operating in states that prohibited branching altogether (UNIT = 0 otherwise). For all quarters 1999.Q1 through 2006.Q4, we set STATEWIDE = LIMITED = UNIT = 0 to reflect the interstate branching regime.

As noted previously, we estimate our empirical model using two alternative samples. The first consists of quarterly observations for 1984.Q1 through 2006.Q4 on all U.S. commercial banks. We omitted banks with missing or negative values for any input or output, and converted dollar values to constant year-2000 prices using the GDP deflator. After pooling the data across quarters, 887,369 observations are available for estimation, with from 5,922 to 13,709 observations in each quarter. The second sample is identical to the first sample for 1984.Q1 through 2000.Q4. For 2001.Q1 through 2006.Q4, however, the second sample consists of quarterly observations on commercial banks that were independent or owned by a non-reporting holding company, and of observations on reporting top-tier holding companies. As with the first sample, we omitted banking organizations with missing or negative values for any input or output, and converted dollar values to constant year-2000 prices. For a top-tier holding companies with missing or invalid data, we included data for its subordinate banks. Table 1 reports summary statistics as of 1984.Q4, 1995.Q4, and 2006.Q4 for total assets and the variables described above. Note that for the second sample we report summary statistics only for 2006.Q4 since the two samples are identical for 1984.Q4 and 1995.Q5 (as well as all other quarters between 1984.Q1 and 2000.Q4).[11]

The distribution of total assets (and other measures of size) among U.S. banks is extremely wide and skewed. Figure 1 shows kernel density estimates for (inflation-adjusted) total assets in 1984.Q4, 1995.Q4, and 2006.Q4 for sample #1, comprised only of commercial banks. The densities for each period are noticeably skewed to the right, despite the use of a log scale on the figure's horizontal axis. The density estimates also reveal that the distribution of bank sizes has shifted to the right, reflecting the increase in mean (and median) bank size over time.[12]

[11] Data for commercial banks are from Reports of Income and Condition ("call reports"), and those for bank holding companies from the FR Y-9C reports. Both are available from the Federal Reserve Bank of Chicago (http://www.chicagofed.org).

[12] Kernel density estimates for sample #2, comprised of both commercial banks and bank holding companies, reveals similar patterns.

The variables defined above suggest the following mapping:

$$\begin{aligned}\big(&\text{ Y1, Y2, Y3, Y4, Z1, Z2, W1/W3, W2/W3, OFF, TIME,}\\ &\text{MBHC, STATEWIDE, LIMITED, UNIT}\big) \rightarrow \text{COST/W3.}\end{aligned} \tag{2.1}$$

Note that we divide COST, W1, and W2 by the price of labor services, W3, to maintain homogeneity with respect to input prices.

This mapping in turn suggests the following regression function:

$$\frac{\text{COST}}{\text{W3}} = C(\boldsymbol{y}, \boldsymbol{w}) + \varepsilon \tag{2.2}$$

where $\boldsymbol{y} = \big[\text{Y1} \quad \text{Y2} \quad \text{Y3} \quad \text{Y4} \quad \text{OFF}\big]$,

$$\boldsymbol{w} = \left[\tfrac{\text{W1}}{\text{W3}} \quad \tfrac{\text{W2}}{\text{W3}} \quad \text{Z1} \quad \text{Z2} \quad \text{TIME} \quad \text{MBHC} \quad \text{STATEWIDE} \quad \text{LIMITED} \quad \text{UNIT}\right],$$

and ε is a random error term with $E(\varepsilon) = 0$ and $\text{VAR}(\varepsilon) = \sigma^2(\boldsymbol{y}, \boldsymbol{w})$. Given that the expectation of ε equals 0, $C(\boldsymbol{y}, \boldsymbol{w}) = E(\text{COST} \mid \boldsymbol{y}, \boldsymbol{w})$ is a conditional mean function that can be estimated by various regression techniques.

2.2 Ray-Scale Economies

Now consider a particular point $(\boldsymbol{y}_0, \boldsymbol{w}_0)$ in the space of $(\boldsymbol{y}, \boldsymbol{w})$. The set of points $\mathfrak{R}_0 = \{(\theta \boldsymbol{y}_0, \boldsymbol{w}_0) \mid \theta \in (0, \infty)\}$ comprises a ray along which the outputs Y1, Y2, Y3, Y4, and OFF are produced in constant proportion to one another. Ray scale economies can be evaluated by examining how expected cost varies along this ray, providing insight into returns to scale along the ray $\mathfrak{R}_0$. Returns to scale are frequently measured in terms of elasticities; the elasticity of cost (with respect to $\boldsymbol{y}$) at a particular point $(\boldsymbol{y}, \boldsymbol{w})$ along the ray $\mathfrak{R}_0$ is given by

$$\eta(\boldsymbol{y}, \boldsymbol{w}) \equiv \left.\frac{\partial \log C(\theta \boldsymbol{y}, \boldsymbol{w})}{\partial \log \theta}\right|_{\theta=1} = \sum_j \frac{\partial \log C(\boldsymbol{y}, \boldsymbol{w})}{\partial \log y_j}, \tag{2.3}$$

where j indexes the elements of $\boldsymbol{y}$. The elasticity in (2.3) is the multi-product analog of marginal cost divided by average cost on the ray $\mathfrak{R}_0$, with $\eta(\boldsymbol{y}, \boldsymbol{w})(<, =, >)1$ implying (increasing, constant, decreasing) returns to scale as outputs in $\boldsymbol{y}$ are expanded along the ray $\mathfrak{R}_0$. Banks for which $\eta(\boldsymbol{y}, \boldsymbol{w}) \neq 1$ are not competitively viable; if banks are subject to

the normal rules of competitive behavior, either a smaller or a larger firm could drive a bank with $\eta(\boldsymbol{y}, \boldsymbol{w}) \neq 1$ from a competitive market.

The measure defined in (2.3) requires estimation of derivatives of the cost function. We employ fully non-parametric estimation methods, as discussed below in Section 3. Non-parametric estimates of derivatives of a function are typically noisier than estimates of the function itself.[13] Hence, we define the ratio

$$\mathcal{S}(\theta \mid \boldsymbol{y}_0, \boldsymbol{w}_0) \equiv \frac{C(\theta \boldsymbol{y}_0, \boldsymbol{w}_0)}{\theta C(\boldsymbol{y}_0, \boldsymbol{w}_0)}. \tag{2.4}$$

It is straightforward to show that

$$\frac{\partial \mathcal{S}(\theta \mid \boldsymbol{y}_0, \boldsymbol{w}_0)}{\partial \theta} \lesseqgtr 0 \Longleftrightarrow \eta(\boldsymbol{y}_0, \boldsymbol{w}_0) \lesseqgtr 1; \tag{2.5}$$

i.e., $\mathcal{S}(\theta \mid \boldsymbol{y}_0, \boldsymbol{w}_0)$ is decreasing (constant, increasing) in θ if returns to scale are increasing (constant, decreasing) at $(\theta \boldsymbol{y}_0, \boldsymbol{w}_0)$ along the ray $\mathfrak{R}_0$ passing through $(\boldsymbol{y}_0, \boldsymbol{w}_0)$. In addition, $\mathcal{S}(1 \mid \boldsymbol{y}_0, \boldsymbol{w}_0) = 1$ by definition. Thus, we investigate ray scale economies (RSE) along a ray $\mathfrak{R}_0$ by estimating $C(\boldsymbol{y}_0, \boldsymbol{w}_0)$ and $C(\theta \boldsymbol{y}_0, \boldsymbol{w}_0)$ for various values of θ, and using confidence bands to determine whether $\mathcal{S}(\theta \mid \boldsymbol{y}_0, \boldsymbol{w}_0)$ is downward or upward sloping.

2.3 Expansion-Path Scale Economies

In the empirical analysis below, we define the fixed point $(\boldsymbol{y}_0, \boldsymbol{w}_0)$ by taking medians of the variables in our model. Of course, few if any banks may be located along the ray $\mathfrak{R}_0$. Although RSE is a convenient measure of scale economies, it may be misleading if most banks are located "far" from $\mathfrak{R}_0$. As an alternative to RSE, we also consider scale economies along each bank's expansion path, i.e., along the path which holds each bank's output mix constant. Consider a bank operating at the point $(\boldsymbol{y}_0, \boldsymbol{w}_0)$, with cost $C(\boldsymbol{y}_0, \boldsymbol{w}_0)$. Let γ be a small positive number, say 0.05, and consider how cost changes as we move from $((1-\gamma)\boldsymbol{y}_0, \boldsymbol{w}_0)$ to $((1+\gamma)\boldsymbol{y}_0, \boldsymbol{w}_0)$; along this path, the output mix remains constant in the sense that relative proportions are maintained. Now let $\theta(1-\gamma)\boldsymbol{y}_0 = (1+\gamma)\boldsymbol{y}_0$; then $\theta = (1+\gamma)/(1-\gamma) \approx 1.1053$ for $\gamma = 0.05$.

[13] This is particularly true for the present case where we would require derivatives in several dimensions; in addition, bandwidth selection becomes problematic when estimating derivatives in more than one dimension.

The following expression provides a measure of expansion-path scale economies (EPSE) for a bank operating at $(\boldsymbol{y}_0, \boldsymbol{w}_0)$:

$$\mathcal{E}_0 = \frac{C\left(\theta(1-\gamma)\boldsymbol{y}_0, \boldsymbol{w}_0\right)}{\theta C\left((1-\gamma)\boldsymbol{y}_0, \boldsymbol{w}_0\right)} \tag{2.6}$$

$$= \frac{C\left((1+\gamma)\boldsymbol{y}_0, \boldsymbol{w}_0\right)}{\left(\frac{1+\gamma}{1-\gamma}\right) C\left((1-\gamma)\boldsymbol{y}_0, \boldsymbol{w}_0\right)}. \tag{2.7}$$

By construction, an increase in output quantities by the factor $\theta > 1$ is associated with an increase in cost by a factor $\mathcal{E}_0\theta$; alternatively, a decrease in outputs by a factor θ^{-1} leads to a decrease in costs by a factor $(\mathcal{E}_0\theta)^{-1}$. Therefore, a bank operating at $(\boldsymbol{y}_0, \boldsymbol{w}_0)$ experiences (decreasing, constant, increasing) returns to scale along the path from $((1-\gamma)\boldsymbol{y}_0, \boldsymbol{w}_0)$ to $((1+\gamma)\boldsymbol{y}_0, \boldsymbol{w}_0)$ as $\mathcal{E}_0(>, =, <)1$. Our measure $\mathcal{E}_0$ provides an indication of returns to scale faced by a particular bank along the path from the origin through the bank's *observed* output vector, starting at a level equal to 95-percent of the quantities in $\boldsymbol{y}_0$ and continuing to a level equal to 105-percent of the quantities in $\boldsymbol{y}_0$ for $\gamma = 0.05$.

The RSE and EPSE measures are both defined in terms of a bank's cost function. The following section discusses a strategy for estimating the cost function non-parametrically, which in turn allows us to estimate, and make inference about, these measures of scale economies.

3 Strategy for Estimation and Inference

Various approaches exist for estimating regression functions (i.e., conditional mean functions) such as the one defined above in (2.2). A common approach is to estimate the function parametrically using a translog specification. However, because the translog function is merely a quadratic specification in log-space, this approach limits the variety of shapes the cost function is permitted to take. Further, because the translog is derived from a Taylor expansion of the cost function around the mean of the data, it makes little sense to use a translog specification to attempt inference about returns to scale over units of widely varying size.

We tested and rejected as a mis-specification the translog functional form for the bank cost relationship. Specifically, for each of the 92 quarters represented by the data in sample #1 and the 24 quarters from 2001.Q1 to 2006.Q4 represented in sample #2 (recall that data

in sample #2 for 2000.Q4 and earlier are identical to corresponding data in sample #1), the four binary dummy variables defined in Section 2 divide the data for a given quarter into as many as eight sub-samples. For each sub-sample in each quarter, we sorted observations by total assets and then split the subsample into two halves (consisting of banks smaller than the median bank in the sub-sample, and those that are larger than the median bank in the sub-sample). We next estimated translog cost functions separately on the two halves of the sub-sample as well as on the full sub-sample, and then performed a Wald test to test whether parameter estimates are stable across the two halves. Among 374 cases, we obtained values for the Wald statistic ranging from 74.15 to 948.8, and corresponding p-values ranging from 0.044 to 3.96×10^{-163}. The largest p-value allows us to reject the translog specification at 5 percent significance; the next-largest p-value was 2.49×10^{-5}. Hence, in every case our data reject the translog specification at any reasonable level of significance.[14]

Rejection of the translog functional form is hardly surprising. Several studies have noted that the parameters of a translog function are unlikely to be stable when the function is fit globally across units of widely varying size.[15] The problem points to the use of non-parametric estimation methods. Although non-parametric methods are less efficient than parametric methods in a statistical sense when the *true* functional form is known, non-parametric estimation avoids the risk of specification error when the true functional form is unknown, as in the present application.

We use fully non-parametric, local-linear and local-quadratic estimators augmented along the lines of Li and Racine (2004), Wilson and Carey (2004) and Wheelock and Wilson (2011) to handle discrete covariates. Both the local-linear and local-quadratic estimators are examples of local order-p polynomial estimators, as is the Nadaraya-Watson kernel regression estimator (Nadaraya, 1964; Watson, 1964). For a locally-fit polynomial of order p used to estimate a derivative of order v, going from $(p-v)$ even to $(p-v)$ odd results in a reduction of bias with no increase in variance (e.g., see Fan and Gijbels (1996) for discussion). Hence,

[14] Additional details and discussion are provided in a separate Appendix A, which is available from the authors on request.

[15] See, for example, Guilkey et al. (1983) and Chalfant and Gallant (1985) for Monte Carlo evidence, and Cooper and McLaren (1996) and Banks et al. (1997) for empirical evidence involving consumer demand. Still others have found a similar problem while estimating cost functions for hospitals (Wilson and Carey, 2004) and for US commercial banks (e.g., McAllister and McManus, 1993; Mitchell and Onvural, 1996; and Wheelock and Wilson, 2001); both hospitals and banks vary widely in terms of size.

we use a local-linear estimator to estimate conditional mean functions (resulting in lower asymptotic mean square error than one would obtain with the Nadaraya-Watson estimator), and a local-quadratic estimator to estimate first derivatives of the conditional mean function.

Non-parametric regression models may be viewed as infinitely parameterized; as such, any parametric regression model (such as the translog cost function) is nested within a non-parametric regression model. Clearly, adding more parameters to a parametric model affords greater flexibility. Non-parametric regression models represent the limiting outcome of adding parameters.[16]

Most non-parametric estimators suffer from the "curse of dimensionality," i.e., convergence rates fall as the number of model dimensions increases. The convergence rate of our estimator would be $n^{1/14}$ with ten continuous explanatory variables, as in (2.2), which is much slower than the usual parametric rate of $n^{1/2}$.[17] The slow convergence rate of our estimator implies that for a given sample size, the order (in probability) of the estimation error we incur with our non-parametric estimator will be larger than the order of the estimation error one would achieve using a parametric estimator in a correctly specified model. Of course, the non-parametric estimation strategy avoids specification error that would likely render meaningless any results that might be obtained using a mis-specified model.[18]

In economic applications, multicollinearity among regressors is often viewed as an annoyance, but here we are able to exploit multicollinearity in our data to reduce the dimensionality of our problem, thereby increasing the convergence rate of our estimators and reducing estimation error. We do this by transforming our continuous regressors to principal components space. Principal components are orthogonal, and eigensystem analysis can be used to determine the information content of each principal component. In particular, we sacrifice a small amount of information (approximately 7.5 percent of the independent linear information in sample #1, and about 7.75 percent of similar information in sample #2) by deleting four principal components, reducing the number of continuous variables in our regression from ten

[16] Several possibilities for non-parametric regression exist. Fan and Gijbels (1996, chapter 1) and Härdle and Linton (1999) provide nice descriptions of non-parametric regression and the surrounding issues.

[17] The convergence rate is unaffected by the inclusion of four binary dummy variables.

[18] Convergence results for non-parametric estimators are often expressed in terms of order of convergence in probability. Briefly, for a sequence (in n) of estimators $\widehat{\zeta}_n$ of some scalar quantity ζ, we can write $\widehat{\zeta} = \zeta + O_p(n^{-a})$ when $\widehat{\zeta}$ converges to ζ at rate n^a, and we say that the estimation error is of order in probability n^{-a}. This means that the sequence of values $n^a|\widehat{\zeta}_n - \zeta|$ is bounded in the limit (as $n \to \infty$) in probability. See Serfling (1980) or Simar and Wilson (2008) for additional discussion.

to six. The transformation to principal components can be inverted, and the interpretation of the estimates is straightforward since our estimator is fully non-parametric.[19]

Use of local-linear and local-quadratic estimators requires that we select two bandwidth parameters to control the smoothing over continuous and discrete dimensions in the data. We use least-squares cross-validation to choose the bandwidths, which amounts to choosing bandwidth values that minimize an estimate of mean integrated square error. In the continuous dimensions, we use a κ-nearest-neighbor bandwidth and a spherically symmetric Epanechnikov kernel function. This means that when we estimate cost at any fixed point of interest in the data space, only the κ observations closest to the fixed point of interest can influence estimated cost at that point. Moreover, among these κ observations, the influence that a particular observation has on estimated cost diminishes with distance from the point at which cost is being estimated. This is the sense in which our estimator is a *local* estimator, and is very different from typical, parametric, *global* estimation strategies (e.g., OLS, maximum likelihood, etc.) where all observations in the sample influence (with equal weight) estimation at any given point in the data space.

For statistical inference about our estimates of returns to scale, we use the wild bootstrap introduced by Härdle (1990) and Härdle and Mammen (1993), which allows us to avoid making specific distributional assumptions. We estimate confidence intervals along the lines used in Wheelock and Wilson (2011). Although our estimators are asymptotically normal, the asymptotic distributions depend on unknown parameters; the bootstrap allows us to avoid the need to estimate these parameters, which would introduce additional noise.[20]

[19] With six continuous dimensions, our estimator converges at rate $n^{1/10}$. To get an idea of the order of estimation error that we face, we can compare the order of estimation error with the size of samples against the same order of estimation error one might obtain with an OLS estimator in a correctly-specified model using a sample of m observations. For example, in sample #1 there are $n = 887,046$ observations; setting $887,046^{-1/10} = m^{-1/2}$ yields $m \approx 15.47$; hence, in our application, we can expect estimation error to be of the same order (in probability) that one would expect to obtain with an OLS estimator and a correctly specified model using only 15–16 observations. One would perhaps not expect to find estimates of high statistical significance when using OLS with only 15 to 16 observations; however, as will be seen below in Section 4, most of our results are highly significant. Additional details about our principal components transformation and non-parametric estimation strategy are given in a separate Appendix B, available from the authors upon request.

[20] Additional details about our inference methods are given in a separate Appendix B, which is available from the authors upon request.

4 Empirical Results

4.1 General Remarks

Recall that samples #1 and #2 differ only over the period 2001.Q1–2006.Q4. We present results for three periods, namely 1984.Q4, 1995.Q4, and 2006.Q4, although we use all observations in a sample for estimation. As a consequence of the local nature of our estimators, the estimates for 1984.Q4 are virtually identical across the two samples, and nearly so for 1995.Q4. Any meaningful differences in estimates between the two samples are possible only for 2006.Q4. In the discussion that follows, we present results obtained using sample #1 for each of the three periods, as well as results obtained using sample #2 for the last period. Results obtained from sample #1 permit comparison across time with a homogeneous sample, while comparisons across samples in the last period serve as a robustness check.

As discussed in Section 3, both our non-parametric local-linear and local-quadratic estimators involve two bandwidth parameters, with one controlling the degree of smoothing along continuous data dimensions, and the other controlling the degree of smoothing across the eight categories defined by our four binary dummy variables. The principal components transformation that we use to reduce dimensionality orthogonalizes the data and standardizes the variances, so that we can use a single bandwidth in the continuous dimensions. For the continuous dimensions, we use nearest-neighbor bandwidths as in Wilson and Carey (2004) and Wheelock and Wilson (2011). For sample #1, least-squares cross validation with the local-linear estimator yields a nearest-neighbor bandwidth $\widehat{\kappa} = 6,477$, meaning that the conditional mean function at a given point is estimated by giving positive weights to the 6,477 observations closest to the point of interest in terms of Euclidean distance, and zero weight to the remaining observations. For sample #2, the nearest-neighbor bandwidth is optimized at $\widehat{\kappa} = 6,420$, and is only slightly smaller than the bandwidth used for sample #1.[21]

[21] Details on cross validation with the local-linear and local-quadratic estimators are available in a separate Appendix B, which is available from the authors upon request. Using the local-quadratic estimator, the optimized bandwidths are somewhat larger; for sample #1, the optimized nearest-neighbor bandwidth is 48,491, while for sample #2 it is 47,486. For both estimators and both samples, the "optimal" values of the bandwidths for the discrete variables is about 0.97–0.99. Optimization of the bandwidth parameters requires numerous evaluations of the cross-validation function for different values of the bandwidth parameters, and the computation time required for each evaluation is proportional of order $O(n^2)$. With almost one million observations, optimization is a formidable task. One approach would be to optimize the bandwidth

4.2 Estimates of RSE

After optimizing bandwidth parameters, we used sample #1 and the local-linear estimator to estimate the RSE measure defined in (2.4) for $\theta \in \{0.05, 0.10, 0.15, \ldots, 0.95, 1.0, 2.0, 3.0, \ldots, 30.0, 32.0, 34.0, \ldots, 120.0\}$, with $(\boldsymbol{y}_0, \boldsymbol{w}_0)$ given by the medians of each variable, setting $TIME$ equal to 4, 48 or 92 (corresponding to 1984.Q4, 1995.Q4, and 2006.Q4). We repeated this exercise using sample #2 for $TIME = 92$, again corresponding to 2006.Q4. The range of values for θ ensures that we consider RSE throughout the range of bank sizes (measured in terms of total assets) in our sample. For sample #1, $\theta = 0.05$ times median total assets corresponds to the 0.02, 0.03, and 0.03 percentiles of total assets in 1984.Q4, 1995.Q4, and 2006.Q4 (respectively), while $\theta = 120$ times median total assets corresponds to the 99.46, 99.08, and 99.36 percentiles of total assets in the same periods. For sample #2, $\theta = 0.05$ times median total assets corresponds to the 0.02 percentile in 2006.Q4, while $\theta = 120$ times median assets corresponds to the 99.11 percentile for the same period.

Figures 2–3 illustrate the estimation results for RSE for 1984.Q4, 1995.Q4, and 2006.Q4. As noted previously, sample #2 is identical to sample #1 for the period 1984-2000; this and the local nature of our estimator ensures that there are no qualititative differences in results across the two samples for 1984.Q4 and 1995.Q4; hence, for those periods we show only results obtained with sample #1. However, for 2006.Q4 we show results for both samples in the bottom two panels of Figure 3.

Although the binary dummy variables MBHC, STATEWIDE, LIMITED, and UNIT define eight cells in each year, we report results only for cells in which banks are observed in the given year. Table 2 reports the number of observations in each of the eight cells for 1984.Q4, 1995.Q4, and 2006.Q4. Unit banking prevailed in a few states in 1984.Q4 but in neither of the later periods, whereas interstate branching was available to all banks in 2006.Q4. For sample #1, Figures 2–3 show the results for banks that were members of multi-bank holding companies (indicated by MBHC = 1), which tend to be larger than independent banks and those owned by single-bank holding companies; the results for banks that were

parameters using a random subset of the observations in our data and then adjust for the true sample size; however, given the highly skewed nature of our data, this might introduce substantial noise if the size of the sub-sample is small. Instead, we wrote Fortran code employing calls to the Message-Passing Interface (MPI) library (see Gropp et al., 1999 for discussion) to compute the cross-validation function in parallel.

not multi-bank holding company members are qualitatively similar.[22] Results shown for 2006.Q4 based sample #2 are for top-tier bank holding companies and independent banks, and therefore encompass the largest banking organizations (by definition, top-tier holding companies do not belong to another holding company, and hence for these institutions, MBHC = 0).

The solid curves in Figures 2–3 trace out our estimates of RSE corresponding to various values of θ. The dashed curves are upper and lower bounds of 99-percent confidence intervals estimated using the bias-corrected bootstrap method.[23] The point estimates lie outside the estimated confidence intervals in some cases because of the bias-correction incorporated into our confidence interval estimation. Note that both axes in Figures 2–3 have log scales.

Recalling the discussion in Section 2, a downward slope for the RSE measure as a function of θ indicates increasing returns to scale along the ray from the origin through the median point $(\boldsymbol{y}_0, \boldsymbol{w}_0)$. The results displayed in Figures 2–3 provide clear evidence of increasing returns to scale throughout the size range of banks in each period, regardless of branching regime or sample used for estimation. From the previous discussion in Section 3 regarding sample size and the "equivalent" sample size in terms of estimation error achieved with an OLS estimator in a correctly specified model, we might a priori expect marginally significant results at best. Yet, our results appear quite strong in terms of statistical significance throughout the range of bank sizes.

4.3 Estimates of EPSE

Whereas RSE measures returns to scale along a ray from the origin through the median point $(\boldsymbol{y}_0, \boldsymbol{w}_0)$, EPSE measures returns to scale for individual banks along the ray from the origin through a given bank's observed output vector, starting at a point equal to 95-percent of the quantities in $\boldsymbol{y}_0$ and continuing to a level equal to 105-percent of the quantities in $\boldsymbol{y}_0$. EPSE may provide a more realistic picture of returns to scale faced by actual banks, especially for those whose output vector differs substantially from the median.

We estimated the EPSE measure defined in (2.7) for each bank observed in 1984.Q4,

[22] Results for banks that were not members of multi-bank holding companies are available from the authors upon request.

[23] The confidence intervals shown in Figures 2–3 are point-wise, rather than simultaneous. However, it seems doubtful that confidence bands would be so wide as to lead to a failure to reject constant returns to scale throughout the size-range of banks.

1995.Q4, and 2006.Q4 for both of our samples. We then estimated 99-percent bias-corrected confidence intervals using bootstrap methods described previously. Next, for each period, we counted the number of cases where the estimated confidence intervals are strictly less than 1.0 (indicating increasing returns to scale), strictly greater than 1.0 (indicating decreasing returns to scale), or contain 1.0 (indicating constant returns to scale). The results, which are tallied in Table 3, indicate that in each period, more than 99.7 percent of banks faced increasing returns to scale along their observed expansion paths. We cannot reject constant returns to scale for the few remaining banks. In particular, we find no evidence that any banks—not even the very largest banks—faced decreasing returns to scale.

For 2006.Q4, we reject at 99-percent significance the null hypothesis of constant returns in favor of increasing returns for all but nine of 6,691 banks in sample #1, and for all but 16 of 6,074 banking organizations in sample #2. Of the nine in sample #1 for which we cannot reject constant returns to scale, the largest, and only one in the top quartile, had total assets of $641.1 million. The next largest had total assets of $172.1 million, and the total assets of the remaining banks for which we fail to reject constant returns are $154.2, $124.2, $42.3, $30.0, $27.6, $12.4, and $5.5 million. Of the 16 banks in sample #2 for which we cannot reject constant returns to scale in 2006.Q4, the largest had total assets of $140.4 billion; the next nine had assets ranging from $3.4 to $78.6 billion, and the remaining seven had assets ranging from $5.5 to $641.1 million. Thus, we reject constant returns to scale in favor of increasing returns for the largest institutions, and for all but one institution with total assets of $100 billion or more.

Given that we test the null hypothesis of constant returns in 2006.Q4 6,691 times for sample #1 and 6,090 times for sample #2, and fail to reject in only 25 of 12,781 cases, we believe that our evidence is strong and compelling. One might expect that a statistical test would fail to reject in a few cases, even if the null is false in every case, simply due to sampling variation, noise, estimation error, etc. In fact, one might reasonably expect a test to fail to reject the null in perhaps many more than 25 cases; the fact that our tests reject the null in 99.8 percent of all cases for 2006.Q4 is a strong result, particularly since we reject constant returns among the largest banks in our two samples.

Figures 4–6 plot the EPSE confidence interval estimates based on the first sample for each bank in 1984.Q4, 1995.Q4, and 2006.Q4, respectively, and Figure 7 plots the EPSE

confidence interval estimates based on the second sample for each bank in 2006.Q4. The four panels in each figure show the confidence intervals for banks across asset-size quartiles (Quartile 1 consists of the smallest 25-percent of banks; Quartile 2 consists of the next-smallest 25-percent, etc.). Within each panel, banks are sorted by the upper bound of their estimated 99-percent confidence interval. As the figures illustrate, the confidence intervals for almost all banks lie entirely below 1.0, indicating that we reject constant returns to scale in favor of increasing returns. The results are consistent across time and across asset-size quartiles in that we reject constant returns in favor of increasing returns for nearly all banks in each sample year and in each asset-size quartile. There are a few banks in each quartile for which we are unable to reject constant returns, but none for which we reject constant returns in favor of decreasing returns to scale. Thus, the evidence for both expansion-path and ray-scale economies indicate that most U.S. banks faced increasing returns to scale as recently as 2006, despite more that 20 years of industry consolidation and increasing average bank size. Further, our evidence is consistent with other recent studies that find that even the largest U.S. bank holding companies operate under increasing returns to scale (Hughes et al., 2001; Feng and Serilitis, 2009).

Estimates of returns to scale can be used to estimate the resource costs associated with the imposition of limits on the size of banks. For example, in the following "back of the envelope" calculations, we estimate the additional cost of operating the four largest U.S. bank holding companies (BHCs), which by the end of 2010 each had total assets exceeding $1 trillion, as firms with no more than $1 trillion of assets. For ease of exposition, in the discussion that follows we give figures in terms of 2010 dollars.

As of December 31, 2010, four BHCs had total assets greater than $1 trillion: Bank of America Corporation, with $2.268 trillion; JP Morgan Chase with $2.118 trillion; Citigroup with $1.914 trillion; and Wells Fargo, with $1.258 trillion; together, these four institutions managed $7.558 trillion of assets. The largest institution in sample #2 (Citigroup) had total assets of $1.885 trillion in 2006.Q4. For simplicity and to be conservative, suppose that constant returns to scale prevail beyond asset size of $1.885 trillion. Then the assets of the four largest BHCs in 2010 could be managed by $7.558/1.885 \approx 4$ organizations without increasing total costs; this would amount to shifting assets among the four largest BHCs in 2010 so that each is the size of Citigroup in 2006.Q4.

Now consider splitting the assets of these hypothetical, new institutions among eight institutions of equal size with total assets of $\$7,558/8 = \$944.75$ billion dollars. How much would costs of managing these assets increase due to this split? To answer this, divide Citigroup's total assets in 2006.Q4 by the asset size of the eight new, hypothetical institutions to obtain $1.885/(7.558/8) \approx 1.9952$. The inverse of this, $1/1.952 \approx 0.5012$, is the factor by which each of the eight new, hypothetical firms would have to be reduced in order to have assets of \$944.75 billion dollars. The average EPSE estimate among the 20 largest institutions in sample #2 in 2006.Q4 is 0.9754. Recalling from the definition of the EPSE measure in (2.7) that $\gamma = 0.5$ and hence $\theta = 1.95/0.95 \approx 1.1053$, and that reducing Citigroup's size by a factor of $\theta^{-\delta} = 0.5012$ would reduce its costs by a factor $(\mathcal{E}\theta)^{-\delta}$. Solving for $\delta = 6.9019$ and setting $\mathcal{E} = 0.9754$, we have $(\mathcal{E}\theta)^{-\delta} \approx 0.5952$. Citigroup's observed cost in 2006 was \$103.8 billion dollars; therefore, assuming outputs are proportional to size, reducing Citigroup's size by a factor of 0.5012 can be expected to reduce its costs by a factor of (1-0.5952).

Four institutions the size of Citigroup in 2006.Q4 might incurs costs of $\$103.8 \times 4 = \$415.2$ billion dollars to manage \$7.558 trillion dollars of assets. If these assets were instead managed by eight institutions with assets of \$944.75 billion dollars, we would expect the costs to be about $\$103.8 \times 8 \times 0.5952 \approx \$494.3$ billion dollars. Hence, a size cap of \$1 trillion dollars could be expected to increase the cost of managing the assets of the four institutions with total assets exceeding \$1 trillion in 2010 by about $494.3 - 415.2$ or \$79.1 billion dollars per year in 2010 dollars.

For comparison, the combined net income of the four largest U.S. bank holding companies in 2003, 2004, 2005, and 2006 were \$41.6 billion, \$42.6 billion, \$57.3 billion, and \$65.5 billion, respectively, in current-year dollars, or \$48.9 billion, \$48.7 billion, \$63.4 billion, and \$70.2 billion in 2010 dollars. Hence, our back of the envelope estimate suggests that capping the size of the four largest bank holding companies at \$1 trillion would result in an increase in the total cost of operating those firms that would exceed their combined profits in each of the four years 2003–06.

Ostensibly, discussions about limiting the size of banks are aimed at limiting the cost of future bailouts. Bailouts are not necessary every year, or even every decade, however, while the cost of about \$79.1 billion dollars per year would be on-going, year after year. In addition, as other institutions grow and reach the \$1 trillion dollar boundary, additional

opportunity costs would be incurred by the size limits. There may well be reasons to limit the size of banks, and a complete analysis of the potential benefits and costs of such a policy is beyond the scope of this paper, but the evidence on scale economies suggests that the likely resource costs of a hard limit on the size of banks are probably not trivial.

4.4 Comparison with Prior Results

Although the results obtained in this paper indicate increasing returns to scale throughout the size distribution of banks, Wheelock and Wilson (2001), by contrast, could not reject constant returns to scale for banks larger than about $500 million of total assets. The present paper differs from Wheelock and Wilson (2001) in three ways. First, the present paper arguably uses a more realistic model of bank costs, borrowed from Berger and Mester (2003). In particular, our model includes measures of off-balance sheet output (OFF) and equity capital (Z2), which were not included in the specification of Wheelock and Wilson (2001). Since banks incur costs in generating non-interest income (our measure of off-balance sheet activity), and larger banks tend to generate proportionately more income from off-balance sheet activities than small banks, failure to include off-balance sheet activity as an output would tend to bias against finding increasing returns to scale.[24] Similarly, failure to control for the level of equity capital would also tend to bias estimates of returns to scale downward since larger banks tend to operate with lower equity ratios, and thus are more leveraged and incur more interest expense than small banks. Hence, the strong result obtained in the present paper may be due, at least in part, to the inclusion of off-balance sheet output and equity capital in the model.

A second difference between the present and earlier paper is that the sample sizes of the present paper paper are roughly 100 times the size of the individual cross sections used by Wheelock and Wilson (2001). Although we use a similar local linear estimator to estimate returns to scale, the larger datasets could also explain why our results differ from those of Wheelock and Wilson (2001).

Finally, a third difference between the two papers lies in the time periods examined; Wheelock and Wilson (2001) used three cross-sectional datasets covering 1985, 1989, and

[24] DeYoung and Roland (2001) show that banks with high levels of non-interest income have disproportionately large amounts of labor expense.

1994. In the present paper, we explicitly incorporate time into the model, which allows us to pool data across time; here, we use data covering 1984.Q1–2006.Q4, and present results for the fourth quarters of 1984, 1995, and 2006. Consequently, while the data used here cover the periods examined in the earlier paper, they also span a dozen years after the last year examined in the earlier paper.

Although the data used in the present paper includes more recent years than those used in Wheelock and Wilson (2001), our non-parametric, local estimation techniques ensure that the estimates for 1984.Q4 and 1995.Q4 are only minimally influenced by data lying far away in time from these periods. In general, the bandwidths used in our local estimators are decreasing in sample size, meaning that as sample size increases, the estimates become increasingly "local." Increasing the amount of data allows increasingly accurate estimates, as less smoothing is required as sample sizes increase. Consequently, the estimates for 1984.Q4 and 1995.Q4 in the present paper are comparable with those for 1984 and 1994 obtained in our earlier paper, and any differences in the results between the two papers for those periods can only be due to differences in model specification or sample size, and not to the inclusion of data from more recent years in the present paper.

To investigate why our results differ from those of Wheelock and Wilson (2001), we estimated the model of bank cost of Wheelock and Wilson (2001) using data pooled over the same 92 quarters that we used to obtain results for the present paper, totaling 885,985 observations. In doing so, we obtained estimates of ray-scale economies that are similar to those reported in Wheelock and Wilson (2001), i.e., increasing returns up to approximately the median bank size, and constant returns to scale for larger banks. However, we obtained estimates of expansion-path scale economies that differ from those reported in the earlier paper, i.e., unlike Wheelock and Wilson (2001) we reject constant returns in favor of increasing returns to scale throughout the range of bank sizes observed in the data. Thus, it appears that including off-balance sheet output and equity capital explains why we obtain evidence of increasing ray-scale economies throughout the range of bank sizes in the present paper when Wheelock and Wilson (2001) did not. By contrast, the finding of increasing expansion-path economies is apparently due to the much larger sample size, and hence greater statistical precision of the estimates reported in the present paper.[25]

[25] Complete estimation results obtained using the variable specification appearing in Wheelock and Wilson

5 Conclusions

Bank executives frequently cite the attainment of scale economies as an important reason for bank mergers and acquisitions, but until recently few studies have found evidence of significant scale economies among banks. Early studies of scale economies in banking typically imposed restrictive parametric specifications or unrealistic assumptions about bank production, however, and more recent studies that use non-parametric estimators or more realistic models of bank production tend to find more evidence of significant scale economies in banking. The present paper adds to a growing body of evidence that banks face increasing returns over a large range of sizes. We use non-parametric local linear estimation to evaluate both ray-scale and expansion-path scale economies for two panel datasets comprised of 1) quarterly observations on all U.S. commercial banks during 1984–2006, and 2) all commercial banks during 1984–2000 and a mix of commercial banks and top-tier bank holding companies during 2001-06. Using either sample, and either measure of scale economies, we find that most U.S. banks operated under increasing returns to scale. The fact that most banks faced increasing returns as recently as 2006 suggests that the U.S. banking industry will continue to consolidate and the average size of U.S. banks is likely to continue to grow unless impeded by regulatory intervention. Thus, our results indicate that while regulatory limits on the size of banks may be justified to ensure competitive markets or to limit the number of institutions deemed too-big-to-fail, such limits could impose significant resource costs on the industry.

(2001) are presented in a separate Appendix C, available from the authors on request.

References

Bach, D. (2001), New megamerger spree seen as tech costs soar, *American Banker* 166, 12.

Banks, J., R. Blundell, and A. Lewbel (1997), Quadratic engel curves and consumer demand, *Review of Economics and Statistics* 79, 527–539.

Berger, A. N. (2003), The economic effects of technological progress: Evidence from the banking industry, *Journal of Money, Credit, and Banking* 35, 141–76.

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

Bernanke, B. S. (2006), Community banking and community bank supervision in the twenty-first century. Remarks at the Independent Community Bankers of America National Convention and Techworld, Las Vegas, Nevada, March 8, 2006.

Chalfant, J. A. and A. R. Gallant (1985), Estimating substitution elasticities with the Fourier cost function, *Journal of Econometrics* 28, 205–222.

Clark, J. A. and T. F. Siems (2002), X-efficiency in banking: Looking beyond the balance sheet, *Journal of Money, Credit and Banking* 34, 987–1013.

Cooper, R. J. and K. R. McLaren (1996), A system of demand equations satisfying effectively global regularity conditions, *Review of Economics and Statistics* 78, 359–364.

DeYoung, R. and K. P. Roland (2001), Product mix and earnings volatility at commercial banks: Evidence from a degree of total leverage, *Journal of Financial Intermediation* 10, 54–84.

Fan, J. and I. Gijbels (1996), *Local Polynomial Modelling and Its Applications*, London: Chapman and Hall.

Feng, G. and A. Serilitis (2009), Efficiency, technical change, and returns to scale in large U.S. banks: Panel data evidence from an output distance function satisfying theoretical regularity, *Journal of Banking and Finance* In press.

Gilbert, R. A. and A. M. Zaretsky (2003), Banking antitrust: Are the assumptions still valid?, *Federal Reserve Bank of St. Louis Review* 85, 29–52.

Greenspan, A. (2010), Testimony before the financial crisis inquiry commission.

Gropp, W., E. Lusk, and A. Skjellum (1999), *Using MPI: Portable Parallel Programming with the Message-Passing Interface*, Cambridge, Massachusetts: The MIT Press.

Guilkey, D. K., C. A. K. Lovell, and R. C. Sickles (1983), A comparison of the performance of three flexible functional forms, *International Economic Review* 24, 591–616.

Haldane, A. G. (2010), The $100 billion question. Comments given at the Insitute of Regulation and Risk, Hong Kong.

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

Härdle, W. and O. Linton (1999), Applied nonparametric methods, in D. L. McFadden and R. F. Engle, eds., *Handbook of Econometrics*, volume 4, chapter 38, New York: Elsevier North-Holland, Inc., pp. 2295–2339.

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

Hughes, J. P., L. J. Mester, and C. G. Moon (2001), Are scale economies in banking elusive or illusive? evidence obtained by incorporating capital structure and risk-taking into models of bank production, *Journal of Banking and Finance* 25, 2169–2208.

Li, Q. and J. Racine (2004), Cross-validated local linear nonparametric regression, *Statistica Sinica* 14, 485–512.

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

Mester, L. J. (2005), Optimal industrial structure in banking. Federal Reserve bank of Philadelphia, Research Department working paper no. 08-2.

Mitchell, K. and N. M. Onvural (1996), Economies of scale and scope at large commercial banks: Evidence from the Fourier flexible functional form, *Journal of Money, Credity, and Banking* 28, 178–199.

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

O'Driscoll, G. P. (2009), The problem with 'nationalization', *The Wall Street Journal* http://online.wsj.com/article/SB123535183265845013.html#printMode, February 23.

Petersen, M. A. and R. G. Rajan (2002), Does distance still matter? the information revolution in small business lending, *The Journal of Finance* 57, 2533–2570.

Reich, R. (2008), If they're too big to fail, they're too big period. Unpublished note available at http://robertreich.blogspot.com/2008_10_01_archive.html.

Rieker, M. (2006), What's driving latest deals? (it's not costs), *American Banker* 171, 1–11.

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

Serfling, R. J. (1980), *Approximation Theorems of Mathematical Statistics*, New York: John Wiley & Sons, Inc.

Simar, L. and P. W. Wilson (2008), Statistical inference in nonparametric frontier models: Recent developments and perspectives, in H. Fried, C. A. K. Lovell, and S. Schmidt, eds., *The Measurement of Productive Efficiency*, chapter 4, Oxford: Oxford University Press, 2nd edition, pp. 421–521.

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

Thompson, L. K. (2002), Efficiency gap separates big banks from samll, *American Banker* 167, 7.

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

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

— (2011), Are credit unions too small?, *Review of Economics and Statistics* Forthcoming.

Wilson, P. W. and K. Carey (2004), Nonparametric analysis of returns to scale and product mix among US hospitals, *Journal of Applied Econometrics* 19, 505–524.

Table 1: Summary Statistics

	Min	1st Quartile	Median	3rd Quartile	Max
		—1984.Q4 (Sample #1 and #2)—			
COST	93.99	2161.81	4126.24	8429.98	17274239.03
Y1	0.00	2638.17	6161.76	13861.81	12738466.49
Y2	0.00	5247.36	10962.29	24065.57	89144075.59
Y3	0.00	3908.97	9854.42	23034.17	29231312.45
Y4	582.85	12099.15	23908.72	49383.34	64865213.54
Z1	1.46	407.91	936.79	2131.34	2597948.49
Z2	20.62	2528.94	4568.31	8869.52	9099694.97
W1$\times 10^5$	30.53	5682.12	7739.64	9254.80	17900.75
W2$\times 10^5$	22.44	5588.75	6670.61	7481.40	17430.40
W3	1.28	26.70	30.94	37.37	119.93
OFF	0.00	50.82	130.81	357.24	3111568.96
MBHC	0.00	0.00	0.00	1.00	1.00
STATEWIDE	0.00	0.00	0.00	0.00	1.00
LIMITED	0.00	0.00	0.00	1.00	1.00
UNIT	0.00	0.00	0.00	1.00	1.00
ASSETS	2385.79	28245.22	53929.17	110151.83	170312400.20
		—1995.Q4 (Sample #1 and #2)—			
COST	116.79	1567.59	2981.96	6147.14	15314803.70
Y1	0.00	2139.57	4787.28	11578.97	15216271.96
Y2	0.00	4684.24	9458.28	19923.62	136667391.39
Y3	0.00	8583.43	20724.37	49762.05	48976138.89
Y4	565.90	14452.84	27746.88	56247.71	83747158.57
Z1	1.07	418.44	1052.81	2555.14	3690692.36
Z2	140.28	3507.85	6751.00	13557.46	16052287.87
W1$\times 10^5$	32.31	3772.26	4511.63	5067.66	17601.41
W2$\times 10^5$	13.97	2493.89	3113.23	3630.11	11320.45
W3	1.58	31.96	37.02	44.20	119.78
OFF	0.00	54.92	148.94	496.61	5446438.08
MBHC	0.00	0.00	0.00	1.00	1.00
STATEWIDE	0.00	1.00	1.00	1.00	1.00
LIMITED	0.00	0.00	0.00	0.00	1.00
UNIT	0.00	0.00	0.00	0.00	0.00
ASSETS	2721.21	35567.12	68007.42	139228.66	235476970.75

Table 1: Summary Statistics (continued)

	Min	1st Quartile	Median	3rd Quartile	Max
			—2006.Q4 (Sample #1)—		
COST	117.69	1774.61	3733.31	8458.59	37039745.79
Y1	0.00	1490.39	3346.72	7818.04	59579635.04
Y2	0.00	6141.04	13413.33	31195.44	139149471.52
Y3	0.00	18181.46	46799.56	119542.79	151342471.86
Y4	756.18	14885.02	29998.49	64198.65	642794052.09
Z1	0.42	626.25	1813.65	4544.54	5611184.46
Z2	442.78	5619.15	10832.90	22853.01	79437664.91
$W1\times10^5$	33.86	3738.21	4172.87	4517.64	17903.20
$W2\times10^5$	10.29	1453.14	2027.74	2612.40	14965.54
W3	9.39	38.26	44.76	54.42	119.12
OFF	0.00	77.37	245.24	853.99	22005093.62
MBHC	0.00	0.00	0.00	0.00	1.00
STATEWIDE	0.00	0.00	0.00	0.00	0.00
LIMITED	0.00	0.00	0.00	0.00	0.00
UNIT	0.00	0.00	0.00	0.00	0.00
ASSETS	4142.88	49706.01	101125.91	231253.37	994201510.38
			—2006.Q4 (Sample #2)—		
COST	117.69	1706.09	3589.98	8409.83	84468206.73
Y1	0.00	1502.84	3360.20	7845.05	102994682.94
Y2	0.00	6042.65	13111.46	31005.32	281977794.95
Y3	0.00	16960.66	43751.36	115828.77	334606705.60
Y4	756.18	14712.34	29150.81	63306.89	937630439.77
Z1	0.42	601.58	1733.89	4507.96	7963296.24
Z2	442.78	5409.41	10505.92	21938.03	113595448.32
$W1\times10^5$	33.86	3678.77	4124.84	4482.79	17903.20
$W2\times10^5$	30.44	1462.01	2029.34	2611.88	11168.07
W3	9.39	38.57	45.14	55.05	119.12
OFF	0.00	72.10	230.46	832.25	41991465.80
MBHC	0.00	0.00	0.00	0.00	1.00
STATEWIDE	0.00	0.00	0.00	0.00	0.00
LIMITED	0.00	0.00	0.00	0.00	0.00
UNIT	0.00	0.00	0.00	0.00	0.00
ASSETS	4142.88	47402.58	97061.05	223970.68	1533643051.37

NOTE: All variables except binary dummy variables (MBHC, STATEWIDE, LIMITED, and UNIT) are measured in 1,000s of U.S. year-2000 dollars.

Table 2: Numbers of Observations by MBHC membership and Branching Restrictions

Branching	1984.Q4	1995.Q4	2006.Q4	2006.Q4*
MBHC=0				
interstate	0	0	5251	5478
statewide	1037	5961	0	0
limited	3734	428	0	0
unit	4045	0	0	0
MBHC=1				
interstate	0	0	1440	612
statewide	246	2620	0	0
limited	1241	258	0	0
unit	1847	0	0	0

NOTE: Asterisk (*) indicates sample #2.

Table 3: Expansion-Path Scale Economies (99-Percent Significance)

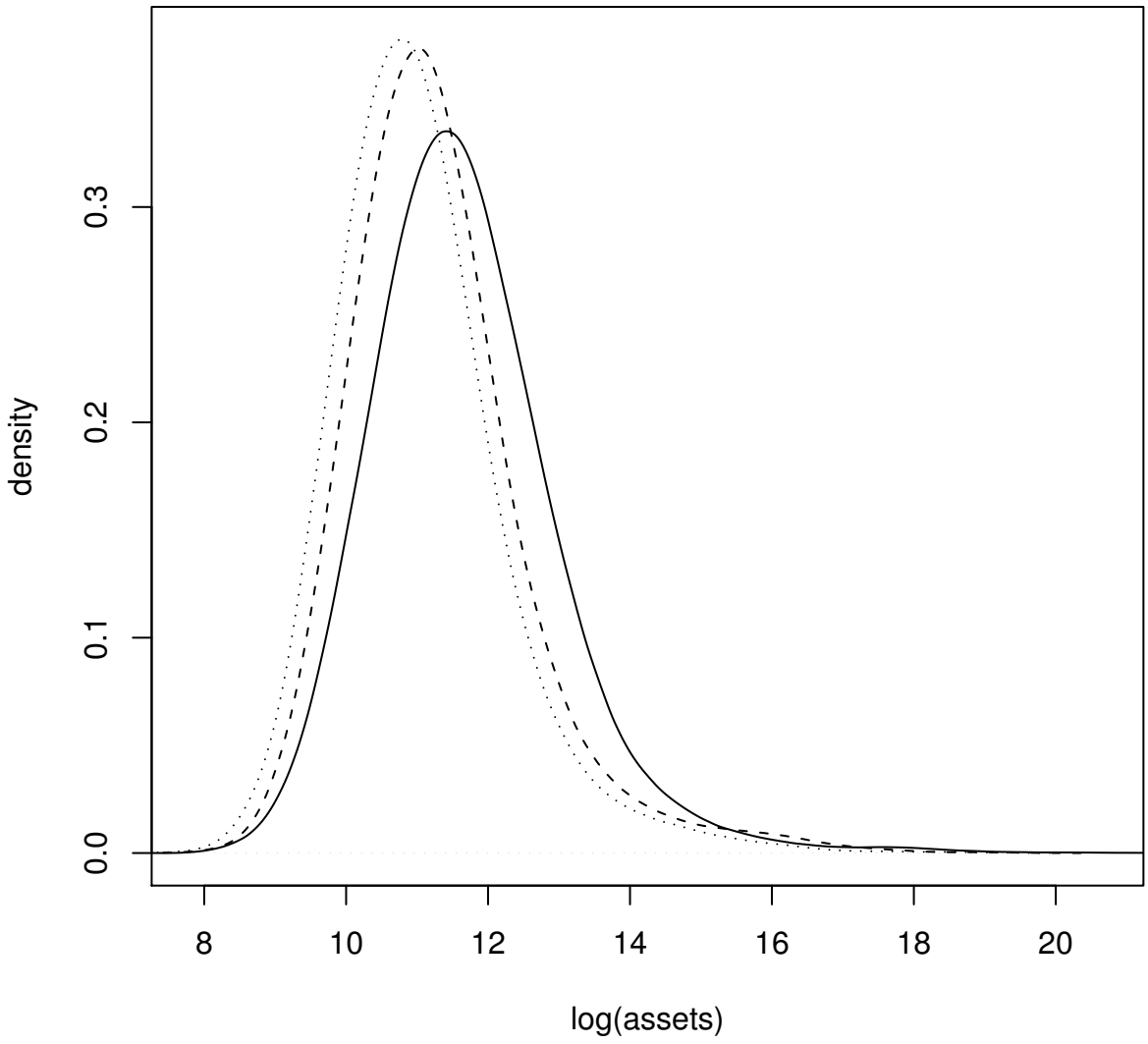
Year	IRS	CRS	DRS
1984	12136	14	0
1995	9241	26	0
2006	6682	9	0
2006*	6074	16	0

NOTE: Asterisk (*) indicates sample #2.

Table 4: Summary Statistics for Expansion-Path Scale Economy Estimates by Size-Quartile

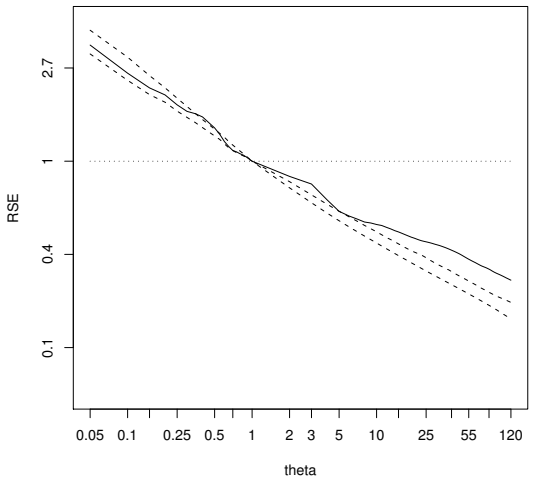
Size Quartile	Min	1st Quartile	Median	Mean	3rd Quartile	Max
			—1984.Q4 (Sample #1)—			
1	0.8935	0.9312	0.9355	0.9360	0.9401	0.9907
2	0.8313	0.9298	0.9354	0.9358	0.9402	1.0059
3	0.8969	0.9313	0.9358	0.9368	0.9410	0.9754
4	0.8855	0.9368	0.9430	0.9436	0.9486	1.0084
			—1995.Q4 (Sample #1)—			
1	0.8789	0.9437	0.9495	0.9454	0.9516	0.9661
2	0.8828	0.9438	0.9498	0.9462	0.9522	0.9685
3	0.8943	0.9452	0.9505	0.9478	0.9532	0.9690
4	0.9098	0.9443	0.9507	0.9496	0.9551	1.0325
			—2006.Q4 (Sample #1)—			
1	0.9308	0.9556	0.9572	0.9562	0.9583	0.9989
2	0.9154	0.9557	0.9573	0.9562	0.9582	0.9739
3	0.9415	0.9560	0.9574	0.9566	0.9584	0.9787
4	0.9103	0.9561	0.9576	0.9577	0.9592	0.9855
			—2006.Q4 (Sample #2)—			
1	0.9322	0.9567	0.9579	0.9576	0.9590	0.9902
2	0.9177	0.9570	0.9582	0.9578	0.9592	0.9662
3	0.9391	0.9574	0.9585	0.9583	0.9595	0.9792
4	0.9429	0.9580	0.9592	0.9602	0.9608	0.9892

Figure 1: Density of (Log) Total Assets for 1984.Q4, 1995.Q4, and 2006.Q4 (Sample #1)

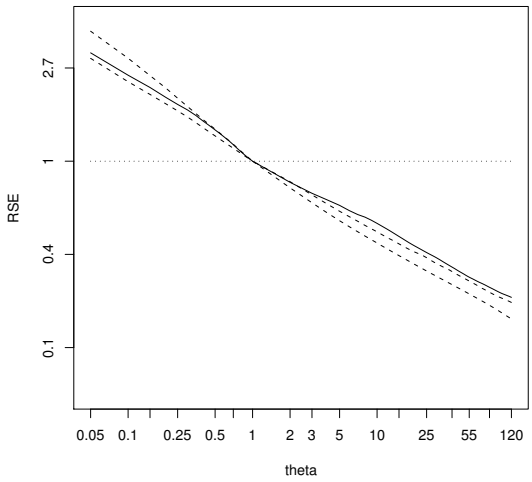


NOTE: The dotted curve gives the density estimate for 1984.Q4; the dashed curve for 1995.Q4, and the solid curve represents 2006.Q4. Total assets are measured in 1,000s of year-2000 U.S. dollars.

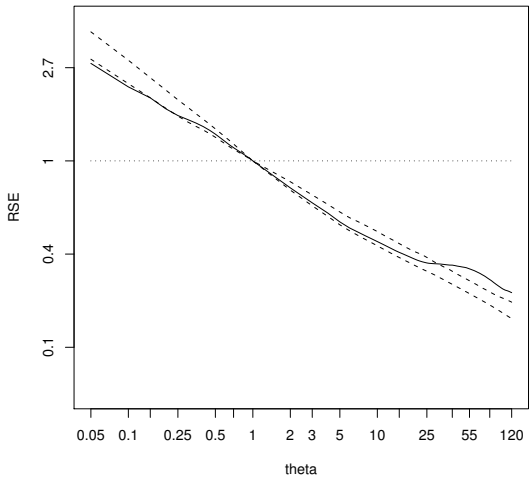
Figure 2: Ray Scale Economies (MBHC = 1, year 1984.Q4, 99-percent significance)



Statewide Branching

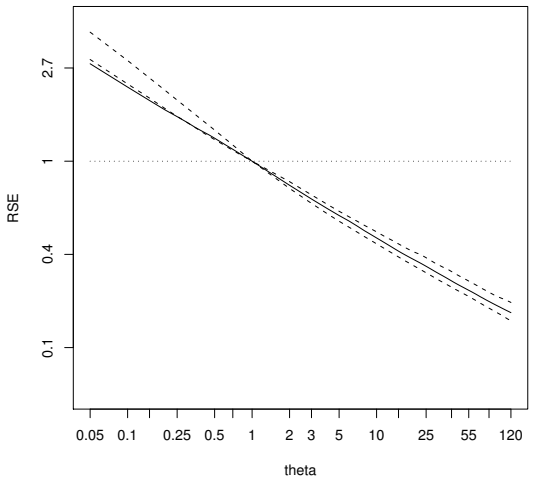


Limited Branching

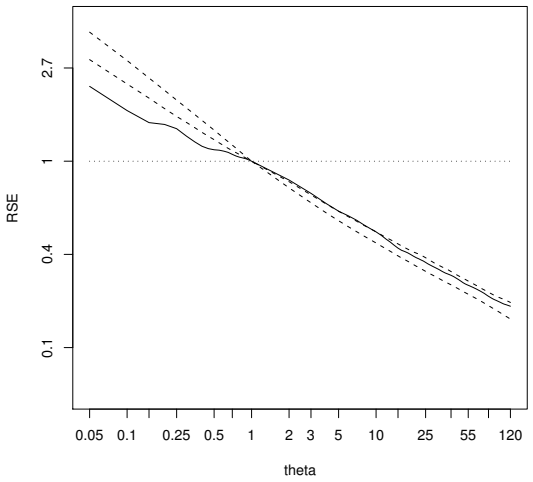


Unit Banking

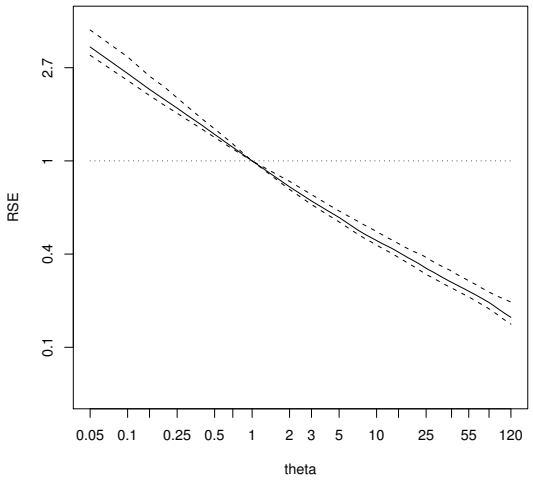
Figure 3: Ray Scale Economies (years 1995.Q4 and 2006.Q5, 99-percent significance)



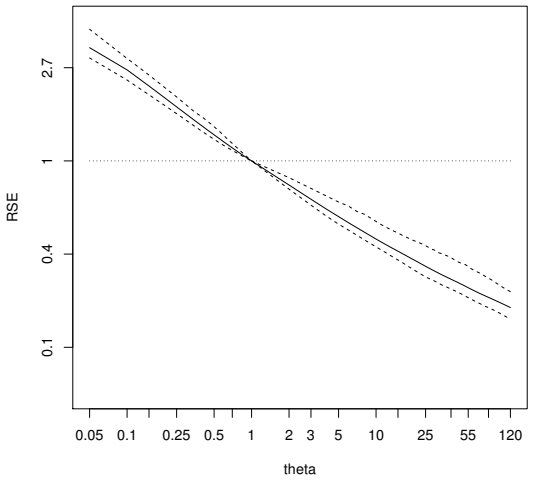
Statewide Branching, 1995.Q4



Limited Branching, 1995.Q4



Interstate Branching, 2006.Q4



Interstate Branching, 2006.Q4*

NOTE: Asterisk (*) indicates results obtained with sample #2. Figures for 1995 and for 2006 using sample #1 correspond to MBHC = 1; Figure for 2006 using sample #2 corresponds to MBHC = 0.

Figure 4: Expansion Path Scale Economies by Size-Quartile, 1984 (Sample #1, 99-percent significance)

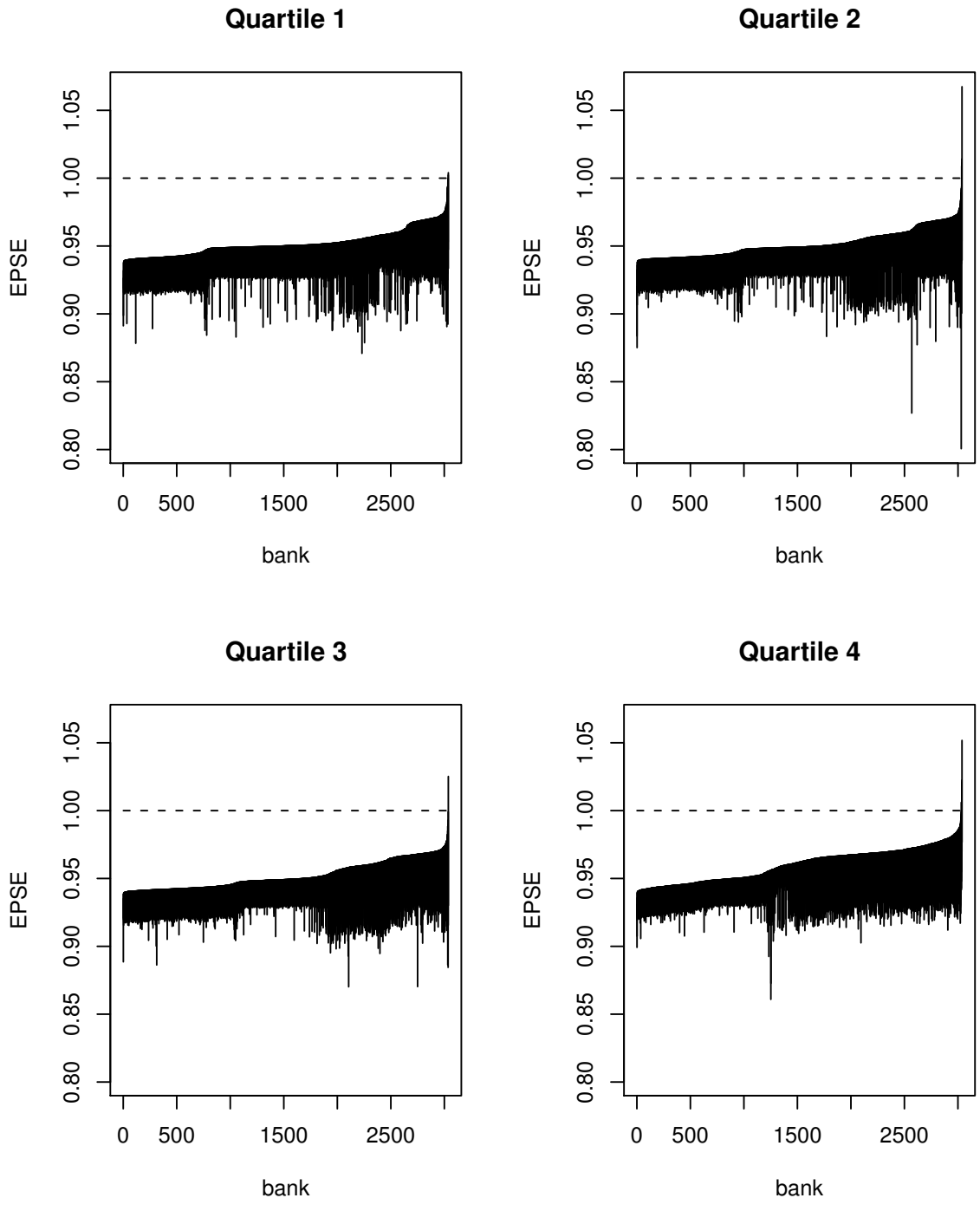


Figure 5: Expansion Path Scale Economies by Size-Quartile, 1995 (Sample #1, 99-percent significance)

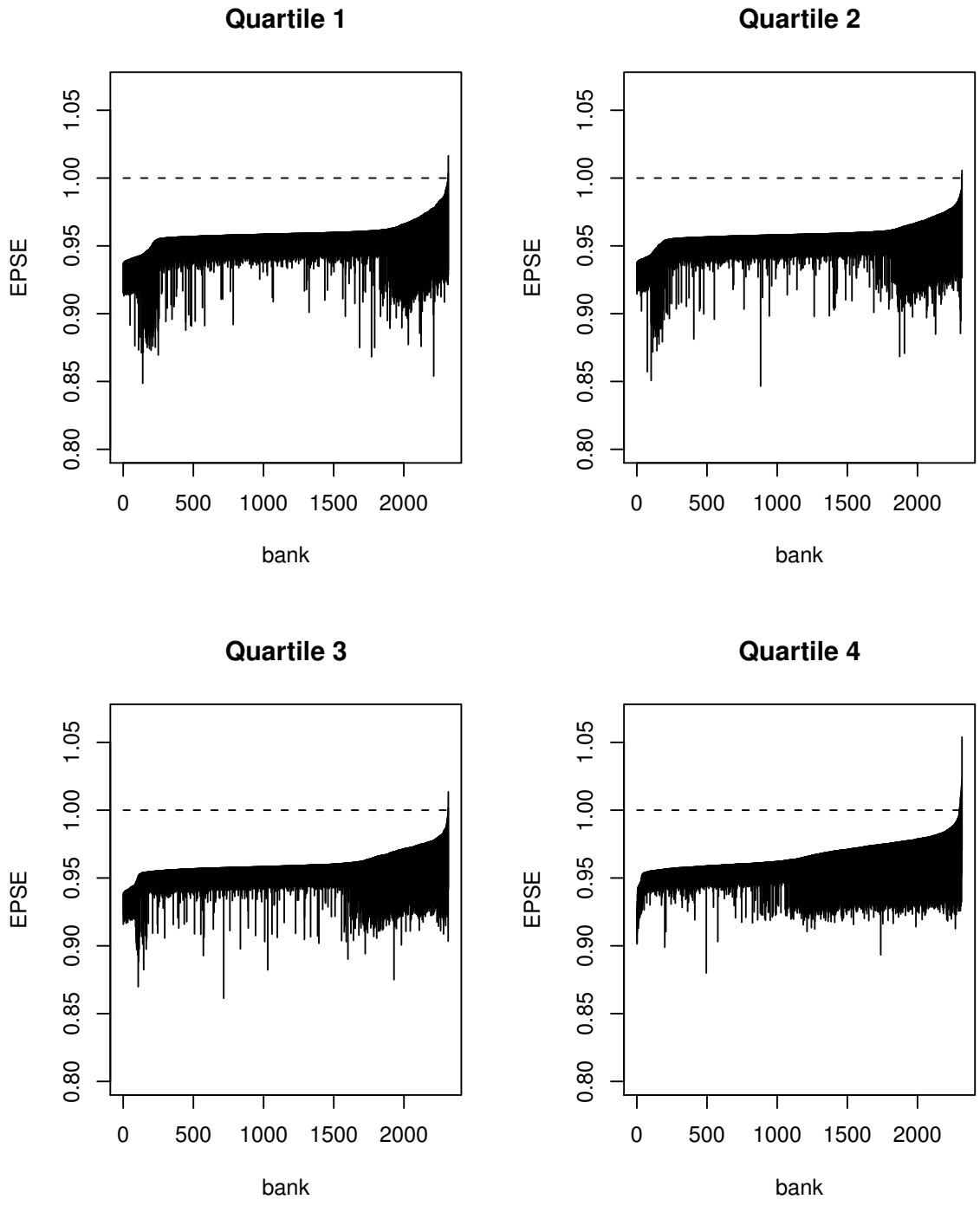


Figure 6: Expansion Path Scale Economies by Size-Quartile, 2006 (Sample #1, 99-percent significance)

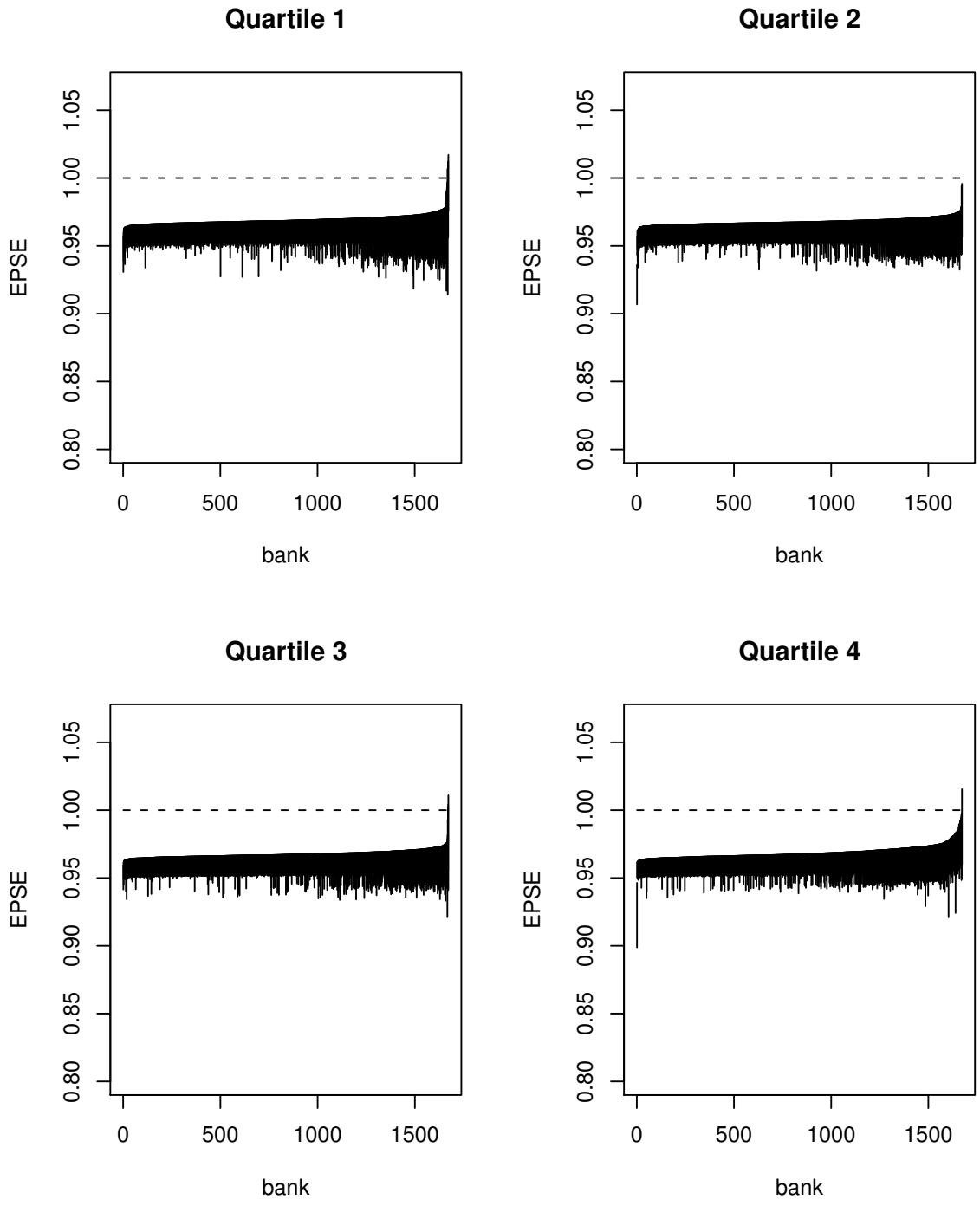
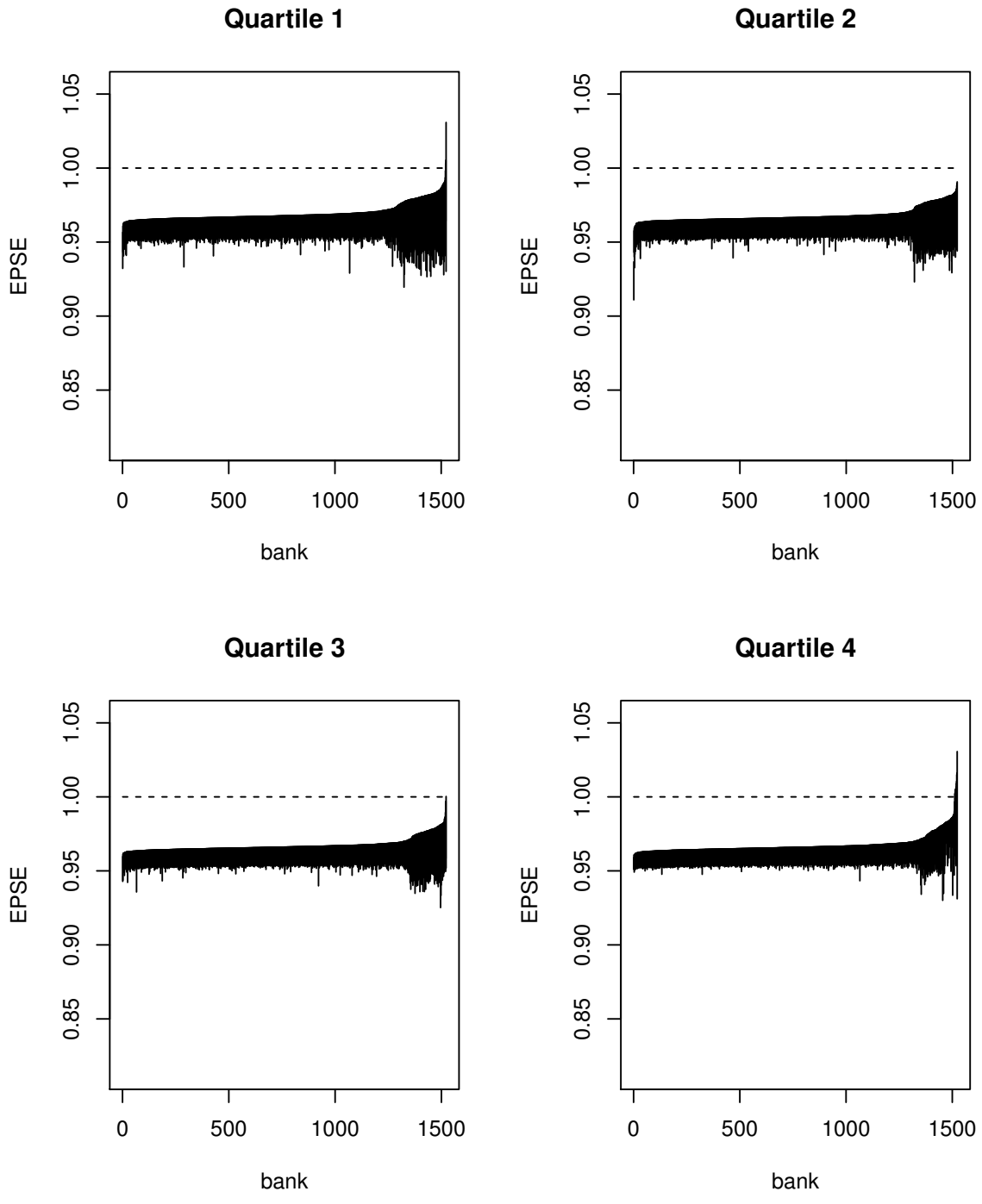


Figure 7: Expansion Path Scale Economies by Size-Quartile, 2006 (Sample #2, 99-percent significance)



Do Large Banks have Lower Costs? New Estimates of Returns to Scale for U.S. Banks Appendices A–D

David C. Wheelock Paul W. Wilson*

May 2011

*Wheelock: Research Department, Federal Reserve Bank of St. Louis, P.O. Box 442, St. Louis, MO 63166–0442; wheelock@stls.frb.org. Wilson: The John E. Walker Department of Economics, 222 Sirrine Hall, Clemson University, Clemson, South Carolina 29634–1309, USA; email pww@clemson.edu. This research was conducted while Wilson was a visiting scholar in the Research Department of the Federal Reserve Bank of St. Louis. We thank the Cyber Infrastructure Technology Integration group at Clemson University for operating the Palmetto cluster used for our computations; we are especially grateful to Barr von Oehsen for technical support and advice. We thank Craig Aubuchon and Heidi Beyer for research assistance. The views expressed in this paper do not necessarily reflect official positions of the Federal Reserve Bank of St. Louis or the Federal Reserve System. *JEL* classification nos.: G21, L11, C12, C13, C14. Keywords: banks, returns to scale, scale economies, non-parametric, regression.

Appendix A: Testing the Translog Functional Form

To test the validity of the translog specification of the bank cost function, we divided our 887,369 sample observations into cells corresponding to unique quarters and unique combinations of the binary dummy variables MBHC, STATEWIDE, LIMITED, and UNIT. With 92 quarters and 8 unique combinations of the binary variables, there are potentially 736 cells; however, some cells are empty (for example, interstate branching—indicated by zero values for STATEWIDE, LIMITED, and UNIT—was prohibited and is hence unobserved in the early years of our sample). For each non-empty cell, we computed the median of total assets and divided the sample in each cell into two sub-samples; for a given cell, sub-sample 1 includes all observations in the cell where total assets within the cell are less than or equal to the cell's median assets, while sub-sample 2 contains all observations within the cell where total assets are greater than the cell's median assets.

Next, for a given cell, we use each of the two subsets to estimate the translog cost model

$$\log(\text{COST}/\text{W3}_i) = \beta_0 + \sum_{j=1}^{9} \beta_j X_{ij} + \sum_{j=1}^{9} \sum_{k=1}^{j} \beta_{jk} X_{ij} X_{ik} + \varepsilon_i \tag{A.1}$$

where $E(\varepsilon) = 0$ and $X_{i\cdot}$ contains the ith observations on the variables log(1+Y1), log(1+Y2), log(1 + Y3), log(Y4), log(Z1), log(1 + Z2), log(W1/W3), log(W2/W3), and log(5 + OFF). The variables COST, W1, and W2 giving dividing variable costs, the price of purchased funds, and the price of core deposits are divided by W3 (price of labor services) to ensure homogeneity with respect to input prices. In addition, it is necessary to add small constants to Y1, Y2, Y3, Z2, and OFF due to small numbers of observations with zero values for these variables.

For sub-sample j containing n_j observations in a given cell, $j \in \{1, 2\}$, let $\boldsymbol{\beta}_j = \begin{bmatrix}\beta_1 & \dots & \beta_{55}\end{bmatrix}'$, and let $\boldsymbol{X}_j$ be the $(n_j \times 55)$ matrix containing the right-hand side variables in (A.1); the first column of $\boldsymbol{X}_j$ consists of a vector of 1s. In addition, let $\boldsymbol{Y}_j$ represent the $(n_j \times 1)$ matrix containing the n_j observations on the left-hand side variable in (A.1), so that the model can be written (for sub-sample j in a given year) as

$$\boldsymbol{Y}_j = \boldsymbol{X}_j \boldsymbol{\beta}_j + \boldsymbol{\varepsilon}_j, \tag{A.2}$$

where $\boldsymbol{\varepsilon}_j$ is an $(n_j \times 1)$ matrix of disturbances with zero mean.

Using data for each sub-sample $j = 1, 2$ in a given cell, we estimate (A.1) using ordinary least squares (OLS), yielding $\widehat{\boldsymbol{\beta}}_j$ and $\widehat{\boldsymbol{\varepsilon}}_j = \boldsymbol{Y}_j - \boldsymbol{X}_j\widehat{\boldsymbol{\beta}}_j$. Next, we compute White's (1980) heteroskedasticity-consistent covariance matrix estimator

$$\widehat{\boldsymbol{\Sigma}}_j = \left(\boldsymbol{X}_j'\boldsymbol{X}_j\right)^{-1}\left(\boldsymbol{X}_j'\text{diag}(\widehat{\boldsymbol{\varepsilon}}_j^2)\boldsymbol{X}_j\right)\left(\boldsymbol{X}_j'\boldsymbol{X}_j\right)^{-1} \tag{A.3}$$

for each sub-sample, where $\text{diag}(\widehat{\boldsymbol{\varepsilon}}_j^2)$ is the $(n_j \times n_j)$ diagonal matrix with elements of $\widehat{\boldsymbol{\varepsilon}}_j^2$ along the principal diagonal. Under the null hypothesis $H_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$, asymptotic normality of OLS estimators ensures that the Wald statistic

$$\widehat{W} = \left(\widehat{\boldsymbol{\beta}}_1 - \widehat{\boldsymbol{\beta}}_2\right)'\left(\widehat{\boldsymbol{\Sigma}}_1 + \widehat{\boldsymbol{\Sigma}}_2\right)^{-1}\left(\widehat{\boldsymbol{\beta}}_1 - \widehat{\boldsymbol{\beta}}_2\right) \xrightarrow{d} \chi^2(55). \tag{A.4}$$

We computed the Wald statistic in (A.4) for each of the 362 non-empty cells represented in our data, obtaining values ranging from 74.15 to 948.8, and corresponding p-values ranging from 0.04356 to 3.963×10^{-163}. The largest p-value allows us to reject the translog specification at 5 percent significance; the next-largest p-value was 2.488×10^{-5}. Hence, in all cases our data reject the translog specification in (A.1) at any reasonable level of significance.

Appendix B: Details of Non-parametric Estimation and Inference

B.1 Dimension reduction

Most non-parametric regression methods suffer from the well-known curse of dimensionality, a phenomenon that causes rates of convergence to become slower, and estimation error to increase dramatically, as the number of continuous right-hand side variables increases (the presence of discrete dummy variables does not affect the convergence rate of our estimator). We use a dimension-reduction technique based on principal components to help mitigate this problem. The idea is to trade a relatively small amount of information in the data for a reduction in dimensionality that will have a large (and favorable) impact on estimation error.

For an $(n \times 1)$ vector $\boldsymbol{V}$ define the function

$$\psi_1(\boldsymbol{V}) \equiv (\boldsymbol{V} - n^{-1}\boldsymbol{i}'\boldsymbol{V})\left[n^{-1}\boldsymbol{V}'\boldsymbol{V} - n^{-2}\boldsymbol{V}'\boldsymbol{i}\boldsymbol{i}'\boldsymbol{V}\right]^{-1/2} \tag{B.1}$$

where $\boldsymbol{i}$ denotes an $(n \times 1)$ vector of 1's. The function $\psi_1(\cdot)$ standardizes a variable by subtracting its sample mean and then dividing by its sample standard deviation. Next, let $\boldsymbol{A}$ be an $(n \times 10)$ matrix with columns $\psi_1(\log(1 + \text{Y1}))$, $\psi_1(\log(1 + \text{Y2}))$, $\psi_1(\log(1 + \text{Y3}))$, $\psi_1(\log(\text{Y4}))$, $\psi_1(\log(\text{Z1}))$, $\psi_1(\log(1 + \text{Z2}))$, $\psi_1(\log(\text{W1/W3}))$, $\psi_1(\log(\text{W2/W3}))$, $\psi_1(\log(5 + \text{OFF}))$, and $\psi_1(\log(\text{TIME}))$.

Let $\boldsymbol{\Lambda}$ be the (10×10) matrix whose columns are the eigenvectors of the correlation matrix of Pearson correlation coefficients for pairs of columns of $\boldsymbol{A}$. Let λ_k be the eigenvalue corresponding to the kth eigenvector in the kth column of $\boldsymbol{\Lambda}$, where the columns of $\boldsymbol{\Lambda}$, and hence the eigenvalues, have been sorted so that $\lambda_1 \geq \ldots \geq \lambda_{10}$. Then set $\boldsymbol{P} = \boldsymbol{A}\boldsymbol{\Lambda}$. The matrix $\boldsymbol{P}$ has dimensions $(n \times 10)$, and its columns are the principal components of $\boldsymbol{A}$. Principal component vectors are orthogonal. Moreover, for each $k \in \{1,\ 2,\ \ldots,\ 10\}$, the quantity

$$\phi_k = \frac{\sum_{j=1}^{k} \lambda_j}{\sum_{\ell=1}^{10} \lambda_\ell} \tag{B.2}$$

represents the proportion of the independent linear information in $\boldsymbol{A}$ that is contained in the first k principal components, i.e., the columns of $\boldsymbol{P}$.

Using our data, we find $\phi_k = 0.5429,\ 0.7699,\ 0.8219,\ 0.8591,\ 0.8930,\ 0.9241,\ 0.9520,\ 0.9756,\ 0.9926$, and 1.0000 for $k = 1,\ \ldots,\ 10$ respectively. Consequently, we use the first six principal components, omitting the last four, in our non-parametric estimation of the bank cost function. In doing so, we sacrifice a relatively small amount of information, while retaining 92.41 percent of the independent linear information in the sample, in order to reduce the dimensionality of our estimation problem by four dimensions in the space of the continuous covariates. This seems a worthwhile trade-off given the curse of dimensionality.

Let $\boldsymbol{P}_{\cdot k}$ denote the kth column of the principal component matrix $\boldsymbol{P}$ and define

$$\psi_0(\boldsymbol{P}_{\cdot k}) \equiv \boldsymbol{P}_{\cdot k} \left[n^{-1} \boldsymbol{P}'_{\cdot k} \boldsymbol{P}_{\cdot k} - n^{-2} \boldsymbol{P}'_{\cdot k} \boldsymbol{i}\boldsymbol{i}' \boldsymbol{P}_{\cdot k} \right]^{-1/2}. \tag{B.3}$$

The transformation $\psi_0(\boldsymbol{P}_{\cdot k})$ has (constant) unit variance. Next, let $\boldsymbol{z}_i$ represent the row vector containing the ith observations on $\psi_0(\boldsymbol{P}_{\cdot 1}),\ \ldots, \psi_0(\boldsymbol{P}_{\cdot 6})$. In addition, let $\boldsymbol{u}_i$ represent the row vector containing the ith observations on the binary variables MBHC, STATEWIDE, LIMITED, and UNIT. We can now write our model as the following regression equation:

$$\mathcal{C}_i = m(\boldsymbol{z}_i, \boldsymbol{u}_i) + \xi_i \tag{B.4}$$

where the subscript i indexes observations, $\mathcal{C}_i = \psi_1(\log(\text{COST/W3}))$, ξ_i is a random error term with $E(\xi_i) = 0$, and $\text{VAR}(\xi_i) = \sigma^2(\boldsymbol{z}_i, \boldsymbol{u}_i)$. The function $m(\boldsymbol{z}_i, \boldsymbol{u}_i) = E(\mathcal{C}_i \mid \boldsymbol{z}_i, \boldsymbol{u}_i)$ is a conditional mean function and can be estimated by non-parametric methods. Moreover, since the transformation from COST to $\mathcal{C}$ can be inverted, given an estimated value $\widehat{m}(\boldsymbol{z}, \boldsymbol{u})$, straightforward algebra leads to an estimate

$$\widehat{C}(\boldsymbol{y}, \boldsymbol{w}) = \exp\left[\psi_1^{-1}\left(\widehat{m}(\boldsymbol{z}, \boldsymbol{u})\right)\right]. \tag{B.5}$$

As discussed below, we use a local linear estimator to estimate $m(\boldsymbol{z}, \boldsymbol{u})$. Although this estimator is weakly consistent, it is asymptotically biased. Moreover, even if $\widehat{m}(\boldsymbol{z}, \boldsymbol{u})$ were *unbiased*, use of the nonlinear transformation in (B.5) means that $\widehat{C}(\boldsymbol{y}, \boldsymbol{w})$ obtained from (B.5) would not, in general, be unbiased because the expectations operator is a linear operator. Furthermore, even if an unbiased estimator of $C(\boldsymbol{y}, \boldsymbol{w})$ were used, plugging the estimator into the definitions of $\mathcal{S}$ and $\mathcal{E}_0$ given in the text to obtain estimators $\widehat{\mathcal{S}}$ and $\widehat{\mathcal{E}}_0$ involves additional nonlinear transformations. Fortunately, any bias in the resulting estimates $\widehat{\mathcal{S}}$ and $\widehat{\mathcal{E}}_0$ can be corrected while making inference about returns to scale; as discussed below in Section B.3, we employ a bias-corrected bootstrap method when estimating confidence intervals for our returns-to-scale measures.

B.2 A non-parametric estimator of the cost relationship

In order to estimate the conditional mean function in (B.4), ignore (for the moment) the time variable T and the binary dummy variables D_1, D_2, so that we can write the conditional mean function on the right-hand side of (B.4) as $m(\boldsymbol{z})$. Both the Nadaraya-Watson (Nadaraya, 1964; Watson, 1964) kernel estimator and the local linear estimator are special cases of local polynomial estimators; with the local linear estimator, the local polynomial is of order 1, while with the Nadaraya-Watson estimator the local polynomial is of order 0. The local linear estimator has less asymptotic bias, but the same asymptotic variance, as the Nadaraya-Watson estimator.

To illustrate the local linear estimator, momentarily ignore the discrete covariates in (B.4) and write the conditional mean function as $m_*(\boldsymbol{z})$. Note that $\boldsymbol{z}$ is a vector of length ℓ. The local linear estimator follows from a first-order Taylor expansion of $m_*(\boldsymbol{z})$ in a neighborhood

of an arbitrary point $\boldsymbol{z}_0$:

$$m_*(\boldsymbol{z}) \approx m_*(\boldsymbol{z}_0) + \frac{\partial m_*(\boldsymbol{z}_0)}{\partial \boldsymbol{z}}(\boldsymbol{z} - \boldsymbol{z}_0). \tag{B.6}$$

This suggests estimating the conditional mean function at $\boldsymbol{z}_0$ by solving the locally weighted least squares regression problem

$$\begin{bmatrix}\widehat{\alpha}_0 & \widehat{\boldsymbol{\alpha}}\end{bmatrix}' = \underset{\alpha_0,\boldsymbol{\alpha}}{\operatorname{argmin}} \sum_{i=1}^{n} \left[\mathcal{C}_i - \alpha_0 - (\boldsymbol{z}_i - \boldsymbol{z}_0)\boldsymbol{\alpha}\right]^2 K\left(|\boldsymbol{H}|^{-1}(\boldsymbol{z}_i - \boldsymbol{z}_0)\right) \tag{B.7}$$

where $K(\cdot)$ is a piece-wise continuous multivariate kernel function satisfying $\int \ldots \int_{\mathbb{R}^\ell} K(\boldsymbol{u})\, d\boldsymbol{u} = 1$ and $K(\boldsymbol{u}) = K(-\boldsymbol{u})$, $u \in \mathbb{R}^\ell$; $\boldsymbol{H}$ is an $\ell \times \ell$ matrix of bandwidths; α_0 is a scalar, and $\boldsymbol{\alpha}$ is an ℓ-vector.

The solution to the least squares problem in (B.7) is

$$\begin{bmatrix}\widehat{\alpha}_0 & \widehat{\boldsymbol{\alpha}}'\end{bmatrix}' = \left(\boldsymbol{Z}'\boldsymbol{\Phi}\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{\Phi}\boldsymbol{\mathcal{C}}, \tag{B.8}$$

where $\boldsymbol{\mathcal{C}} = \begin{bmatrix}\mathcal{C}_1 & \ldots & \mathcal{C}_n\end{bmatrix}'$, $\boldsymbol{\Phi} = \operatorname{diag}\left[K\left(|\boldsymbol{H}|^{-1}(\boldsymbol{z}_i - \boldsymbol{z}_0)\right)\right]$, and $\boldsymbol{Z}$ is an $n \times (\ell + 1)$ matrix with ith row given by $\begin{bmatrix}1 & (\boldsymbol{z}_i - \boldsymbol{z}_0)\end{bmatrix}$. The fitted value $\widehat{\alpha}_0$ provides an estimate $\widehat{m}_*(\boldsymbol{z}_0)$ of the conditional mean function $m_*(\boldsymbol{z}_0)$ at an arbitrary point $\boldsymbol{z}_0$.[1]

Some modifications are necessary to introduce the binary dummy variables D_{i1} and D_{i2} into the analysis. One possibility is to split the sample into four sub-samples based on the values of the discrete variables, and then analyze each group separately while treating time as a continuous variable. However, this approach does not make efficient use of the data because each sub-sample may contain some information that would be useful in estimation on the other sub-samples. In addition, in our application, some of the resulting sub-samples would be very small. With the local linear estimator, we can introduce discrete variables by modifying the weights in the weighting matrix $\boldsymbol{\Phi}$ introduced in (B.8). The idea involves smoothing across the four categories defined by the two binary dummy variables, and then letting the data determine how much smoothing is appropriate. Aitchison and Aitken (1976)

[1] The fitted values in $\widehat{\boldsymbol{\alpha}}$ provide estimates of elements of the vector $\partial m(\boldsymbol{z}_0)/\partial \boldsymbol{z}$. However, if the object is to estimate first derivatives, mean-square error of the estimates can be reduced by locally fitting a quadratic rather than a linear expression (see Fan and Gijbels, 1996 for discussion); this increases computational costs, which are already substantial for the local linear fit. Moreover, determining the optimal bandwidths becomes more difficult and computationally more burdensome for estimation of derivatives. See Härdle (1990, pp. 160–162) for discussion of some of the issues that are involved with bandwidth selection for derivative estimation.

discuss the use of a discrete kernel for discrimination analysis. Bierens (1987) and Delgado and Mora (1995) suggest augmenting the Nadaraya-Watson estimator with a discrete kernel, and prove that the estimator remains consistent and asymptotically normal. Racine and Li (2004) establish convergence rates for the Nadaraya-Watson estimator with mixed continuous-discrete data; the rate with continuous and discrete covariates is the same as the rate with the same number of continuous variables, but no discrete variables. The introduction of discrete covariates does not exacerbate the curse of dimensionality, at least in the limit.

It is straightforward to extend the local linear estimator to accommodate discrete dummy variables. Let $\boldsymbol{u}_i$ represent a vector of observations on $k = 4$ binary dummy variables, and consider an arbitrary Bernoulli vector $\boldsymbol{u}_0$ of length k. Then let $\delta(\boldsymbol{u}_i, \boldsymbol{u}_0) = (\boldsymbol{u}_i - \boldsymbol{u}_0)'(\boldsymbol{u}_i - \boldsymbol{u}_0)$, and define the discrete kernel function

$$G(\boldsymbol{u}_i \mid \boldsymbol{u}_0, h_1) = h_1^{k-\delta(\boldsymbol{u}_i,\boldsymbol{u}_0)}(1-h_1)^{\delta(\boldsymbol{u}_i,\boldsymbol{u}_0)} \tag{B.9}$$

where $h_1 \in [\frac{1}{2}, 1]$ is a bandwidth parameter.

Note that $\lim_{h_1 \to 1} G(\boldsymbol{u}_0 \mid \boldsymbol{u}_i, h_1)$ equals either 1 or 0, depending on whether $\boldsymbol{u}_0 = \boldsymbol{u}_i$ or $\boldsymbol{u}_0 \neq \boldsymbol{u}_i$, respectively. In this case, estimation yields the same results as would be obtained if estimation was performed separately on each of the four sub-samples delineated by the dummy variables. Alternatively, if $h_1 = \frac{1}{2}$, then $G(\boldsymbol{u}_0 \mid \boldsymbol{u}_i, h_1) = 1$ regardless of whether $\boldsymbol{u}_0 = \boldsymbol{u}_i$ or $\boldsymbol{u}_0 \neq \boldsymbol{u}_i$; in this case, there is complete smoothing over the four categories, and including the dummy variables has no effect relative to the case where they are ignored.

We specify the kernel function $K(\cdot)$ as an ℓ-variate spherically symmetric Epanechnikov kernel with a single, scalar bandwidth h_0; i.e.,

$$K(\boldsymbol{t}) = \frac{\ell(\ell+2)}{2S_\ell}(1 - \boldsymbol{t}\boldsymbol{t}')I(\boldsymbol{t}\boldsymbol{t}' \leq 1) \tag{B.10}$$

where $I(\cdot)$ is the indicator function, $S_\ell = 2\pi^{\ell/2}/\Gamma(\ell/2)$, $\Gamma(\cdot)$ denotes the gamma function, $\boldsymbol{u} = |\boldsymbol{H}|^{-\ell}(\boldsymbol{z}_i - \boldsymbol{z}_0)$, and $\boldsymbol{H}$ is an $(\ell \times \ell)$ matrix of bandwidths. The spherically symmetric Epanechnikov kernel is optimal in terms of asymptotic minimax risk; see Fan et al. (1997) for details and a proof.

Let $\mathbb{D} = \{0,1\} \times \{(1,0,0),\ (0,1,0),\ (0,0,1),\ (0,0,0)\}$ be the set of possible values for the vector $\boldsymbol{u}$ of binary variables. Incorporating the discrete covariates, an estimate $\widehat{m}\,(\boldsymbol{z}_0, \boldsymbol{u}_0)$ of

the conditional mean function in (B.4) at an arbitrary point $(\boldsymbol{z}_0, \boldsymbol{u}_0) \in \mathbb{R}^\ell \times \mathbb{D}$ is given by $\widehat{\alpha}_0$ obtained from

$$\begin{bmatrix}\widehat{\alpha}_0 & \widehat{\boldsymbol{\alpha}}\end{bmatrix}' = \underset{\alpha_0, \boldsymbol{\alpha}}{\operatorname{argmin}} \sum_{i=1}^{n} \left[\mathcal{C}_i - \alpha_0 - (\boldsymbol{z}_i - \boldsymbol{z}_0)\boldsymbol{\alpha}\right]^2 K\left(|\boldsymbol{H}|^{-1}(\boldsymbol{z}_i - \boldsymbol{z}_0)\right) G(\boldsymbol{u}_i \mid \boldsymbol{u}_0, h_1) \quad \text{(B.11)}$$

where $\boldsymbol{u}_0 \in \mathbb{D}$. The solution to the least-squares problem in (B.11) is given by

$$\begin{bmatrix}\widehat{\alpha}_0 & \widehat{\boldsymbol{\alpha}}\end{bmatrix}' = \left(\boldsymbol{Z}'\boldsymbol{\Omega}\boldsymbol{Z}\right)^{-1} \boldsymbol{Z}'\boldsymbol{\Omega}\mathcal{C}, \quad \text{(B.12)}$$

where $\boldsymbol{Z}$ is defined as in (B.8) and the matrix $\boldsymbol{\Omega}$ of weights is given by

$$\boldsymbol{\Omega} = \operatorname{diag}\left[K(|\boldsymbol{H}|^{-\ell}(\boldsymbol{z}_i - \boldsymbol{z}_0)G(\boldsymbol{u}_i \mid \boldsymbol{u}_0, h_1)\right]. \quad \text{(B.13)}$$

Finally, recall that the principal components transformation pre-whitens the data; in addition, the principal components are orthogonal. Orthogonality suggests setting off-diagonal elements to zero. The transformation in (B.3) ensures that the columns of $\boldsymbol{Z}$ have constant, unit variance, suggesting use of the same bandwidth in each direction. Hence we set $\boldsymbol{H} = \operatorname{diag}(h(\boldsymbol{z}_0))$ so that $|\boldsymbol{H}|^{-\ell} = h(\boldsymbol{z}_0)^{-\ell}$, where $h(\boldsymbol{z}_0)$ is an adaptive scalar bandwidth depending on the point $\boldsymbol{z}_0$ where the conditional mean function is to be evaluated.

B.3 Practical issues for implementation

To implement our estimator, we must choose values for the bandwidths $h(\boldsymbol{z}_0)$ and h_1. For the discrete variables, we employ a (globally) constant bandwidth, while for the continuous variables we use an adaptive, nearest-neighbor bandwidth. We define $h(\boldsymbol{z}_0)$ for any particular point $\boldsymbol{z}_0 \in \mathbb{R}^\ell$ as the maximum Euclidean distance between $\boldsymbol{z}_0$ and the κ nearest points in the observed sample $\{\boldsymbol{z}_i\}_{i=1}^n$, $\kappa \in \{2,\ 3,\ 4,\ \ldots\}$. Thus, given the data and the point $\boldsymbol{z}_0$, the bandwidth $h(\boldsymbol{z}_0)$ is determined by κ, and varies depending on the density of the continuous explanatory variables locally around the point $\boldsymbol{z}_0 \in \mathbb{R}^\ell$ at which the conditional mean function is estimated. This results in a bandwidth that is increasing with decreasing density of the data around the point of interest, $\boldsymbol{z}_0$. More smoothing is required where data are sparse than where data are dense; our nearest-neighbor bandwidth adapts automatically to the density of the data. The discrete kernels in (B.13) in turn give more (or less) weight to observations among the κ nearest neighbors that are close (or far) away along the time

dimension, or that are in the same (or different) category determined by the combination of binary dummy variables.

Note that we use a nearest-neighbor *bandwidth* rather than a nearest-neighbor *estimator*. The bandwidth is used inside a kernel function, and the kernel function integrates to unity. Loftsgaarden and Quesenberry (1965) use this approach in the density estimation context to avoid nearest-neighbor density estimates (as opposed to bandwidths) that do not integrate to unity (see Pagan and Ullah, 1999, pp. 11-12 for additional discussion). Fan and Gijbels (1994; 1996, pp. 151–152) discuss nearest neighbor bandwidths in the regression context.

As a practical matter, we set $\kappa = [\omega n]$, where $\omega \in (0, 1)$, n represents the sample size, and $[a]$ denotes the integer part of a. We optimize the choice of values for the bandwidth parameters by minimizing the least-squares cross-validation function; i.e., we select values

$$\begin{bmatrix} \widehat{\omega} & \widehat{h}_1 \end{bmatrix}' = \underset{\omega, h_1}{\operatorname{argmin}} \sum_{i=1}^{n} \left[\mathcal{C}_i - \widehat{m}_{-i}(\boldsymbol{z}_i, \boldsymbol{u}_i) \right]^2, \tag{B.14}$$

where $\widehat{m}_{-i}(\boldsymbol{z}_i, \boldsymbol{u}_i)$ is computed the same way as $\widehat{m}(\boldsymbol{z}_i, \boldsymbol{u}_i)$, except that the ith diagonal element of $\boldsymbol{\Psi}$ is replaced with zero. The least-squares cross validation function approximates the part of mean integrated square error that depends on the bandwidths.[2]

Once appropriate values of the bandwidth parameters have been selected, the conditional mean function can be estimated at any point $(\boldsymbol{z}_0, \boldsymbol{u}_0) \in \mathbb{R}^{\ell} \times \mathbb{D}$. We then estimate the RSE and EPSE measures defined in in the text by replacing the cost terms with estimates obtained from the relation (B.5). To make inferences about RSE and EPSE, we use the wild bootstrap proposed by Härdle (1990) and Härdle and Mammen (1993).[3] We obtain bootstrap estimates

[2] Choice of κ by cross validation has been proposed by Fan and Gijbels (1996) and has been used by Wheelock and Wilson (2001) and Wilson and Carey (2004) and others. Using n_p CPUs, the computation time required for each evaluation of the cross-validation function is only slightly more than $1/n_p$ times the time that would be required on a single processor. We performed all computations on the Palmetto cluster operated by Clemson University's Cyber Infrastructure Technology Integration (CITI) group. Our code was run on nodes with dual AMD Opteron 2356, 2.3Ghz processors; each processor has 4 cores, and each node has 16 gigabytes of memory. Hence each node is capable of running 8 threads simultaneously. To optimize the bandwidths for the local-linear estimators using sample #1, we ran our code on 96 quad-core processors, executing 768 threads simultaneously; the optimization required roughly 13.75 hours on each of 96 8-core nodes, or about 10,560 total CPU hours. The local-quadratic estimator requires more time to compute than the local-linear estimator; using sample #1, optimization of the bandwidth parameters for the local-quadratic estimator consumed roughly 221,184 hours of total CPU time. Similar costs were incurred in optimizing bandwidths for sample #2.

[3] Ordinary bootstrap methods are inconsistent in our context due to the asymptotic bias of the estimator; see Mammen (1992) for additional discussion.

$\widehat{m}_b^*(\cdot)$, which we substitute into the definitions of $\mathcal{S}$ and $\mathcal{E}_0$ in the text to obtain bootstrap values $\widehat{\mathcal{S}}_b^*$ and $\widehat{\mathcal{E}}_b^*$ for particular values of $\boldsymbol{z}$ and $\boldsymbol{u}$, with $b = 1, \ldots, B$.

To make inference about $\mathcal{S}$, we use the bias-correction method described by Efron and Tibshirani (1993). In particular, we estimate $(1-\alpha) \times 100$-percent confidence intervals by $\left(\widehat{\mathcal{S}}^{*(\alpha_1)}, \widehat{\mathcal{S}}^{*(\alpha_2)}\right)$, where $\widehat{\mathcal{S}}^{*(\alpha)}$ denotes the α-quantile of the bootstrap values $\widehat{\mathcal{S}}_b^*$, $b = 1, \ldots, B$, and

$$\alpha_1 = \Phi\left(\widehat{\varphi}_0 + \frac{\widehat{\varphi}_0 + \varphi^{(\alpha/2)}}{1 - \widehat{\varphi}_0 + \varphi^{(\alpha/2)}}\right), \tag{B.15}$$

$$\alpha_2 = \Phi\left(\widehat{\varphi}_0 + \frac{\widehat{\varphi}_0 + \varphi^{(1-\alpha/2)}}{1 - \widehat{\varphi}_0 + \varphi^{(1-\alpha/2)}}\right), \tag{B.16}$$

$\Phi(\cdot)$ denotes the standard normal distribution function, $\varphi^{(\alpha)}$ is the $(\alpha \times 100)$-th percentile of the standard normal distribution, and

$$\widehat{\varphi}_0 = \Phi^{-1}\left(\frac{\#\{\widehat{\mathcal{S}}_b^* < \widehat{\mathcal{S}}\}}{B}\right), \tag{B.17}$$

with $\Phi^{-1}(\cdot)$ denoting the standard normal quantile function (e.g., $\Phi^{-1}(0.95) \approx 1.6449$).

For RSE, we sort the values in $\left\{\left(\widehat{\mathcal{S}}_b^* - \widehat{\mathcal{S}}\right)\right\}_{b=1}^{B}$ by algebraic value, delete $(\frac{\alpha}{2} \times 100)$-percent of the elements at either end of this sorted array, and denote the lower and upper endpoints of the remaining, sorted array as $-b_\alpha^*$ and $-a_\alpha^*$, respectively. Then a bootstrap estimate of a $(1-\alpha)$-percent confidence interval for $\mathcal{S}$ is

$$\widehat{\mathcal{S}} + a_\alpha^* \leq \mathcal{S} \leq \widehat{\mathcal{S}} + b_\alpha^*. \tag{B.18}$$

The idea underlying (B.18) is that the empirical distribution of the bootstrap values $\left(\widehat{\mathcal{S}}_b^* - \widehat{\mathcal{S}}\right)$ mimics the unknown distribution of $\left(\widehat{\mathcal{S}} - \mathcal{S}\right)$, with the approximation improving as $n \to \infty$. As $B \to \infty$, the choices of $-b_\alpha^*$ and $-a_\alpha^*$ become increasingly accurate estimates of the percentiles of the distribution of $\left(\widehat{\mathcal{S}}_b^* - \widehat{\mathcal{S}}\right)$ (we set $B = 1000$). Any bias in $\widehat{\mathcal{S}}$ relative to $\mathcal{S}$ is reflected in bias of $\widehat{\mathcal{S}}^*$ relative to $\widehat{\mathcal{S}}$; in the case of large bias, it is conceivable that the estimated confidence interval may not include the original estimate $\widehat{\mathcal{S}}$, since the estimated confidence interval corrects for the bias in $\widehat{\mathcal{S}}$. We estimate confidence intervals for the EPSE measures similarly.

Appendix C: Results Using Variable Specification of Wheelock and Wilson (2001)

In order to explore the possible reasons for differences in results between Wheelock and Wilson (2001) and our current paper, we computed the variables that were used in our earlier paper for each of the 92 quarters 1984.Q1–2006.Q4. This resulted in 885,985 observations available for estimation. The earlier specification contains 9 continuous right-hand side variables; we employed the same dimension-reduction technique described in Appendix B to reduce dimensionality to 5 continuous dimensions, sacrificing 6.72 percent of the independent linear information in the sample. We then estimated the non-parametric cost function defined in our earlier paper using the same local-linear estimator that we use in our current paper, and used these estimates to construct estimates of our RSE and EPSE measures defined in our current paper. The estimation procedure, as well as inference using bootstrap methods, is identical to that used in our current paper; only the variable specifications differ.

Tables C.3–C.4 and Figures C.2–C.6 in this Appendix are analogous to Tables 3–4 and Figures 2–6 in the main part of our current paper. Here as in the main part of our paper, 95-percent significance levels are used.

Appendix D: Do Banks Efficiently Minimize Costs?

The analysis in this paper is based on estimates of the conditional mean function in equation (2.2), rather than estimates of a cost frontier. Because we use non-parametric estimators of the conditional mean function, and due to the ensuing substantial computational burden, it is not feasible to replace the error term in equation (2.2) with a composite error consisting of a two-sided noise term and a one-sided inefficiency process. The mean cost function in equation (2.2) is, however, well-defined and of interest to policy makers. If banks are technically inefficient, then necessarily they operate in the interior of the set of feasible cost-output combinations. The conditional mean function defined in equation (2.2) describes what banks are actually doing, as opposed to what they might do in a perfect world with no inefficiency.

Regardless of whether banks are technically efficient, one can ask whether banks use the optimal amount of capital given their observed levels of outputs and other factors in equation

(2.2). To answer this question, we estimate derivatives of the cost relation in equation (2.2) with respect to financial equity capital (Z2) using a local-quadratic estimator (described in the separate Appendix B, which is available from the authors upon request). Non-parametric estimation of derivatives in multiple regression settings involves a formidable, unsolved question of how to choose bandwidths. Theoretical results discussed in Fan and Gijbels (1996) make clear that bandwidths for derivative estimation must be larger than those used for estimation of the conditional mean function in order to minimize asymptotic mean square error of the derivative estimates. However, to date, there is no tractable, reliable method for optimizing bandwidths for derivative estimation in multiple regression problems.

To proceed, we optimize bandwidths for our local-quadratic estimator of the conditional mean function, and then scale the optimized bandwidths by factors 1, 1.05, 1.1, 1.15, and 1.2 to estimate derivatives. Comparisons of derivative estimates across the five different scaling factors indicates that the results are remarkably robust with respect to the choice of scaling. In the discussion that follows, we present results obtained with the scale factor 1.1.

As noted earlier, optimization of bandwidths for estimation of the conditional mean function with the local-quadratic estimator involves a large computational burden. In addition, computation of the local derivative estimates themselves involves substantial computational burden, and the estimates must be bootstrapped in order to make inference. Consequently, we focused our efforts on the last quarter represented in sample #2, i.e., 2006.Q4. Hughes et al. (2001) suggest that the unobserved price of equity capital likely falls in the interval $[0.14, 0.18]$. Following their approach, we employ one-sided bootstrap tests to test the null that the derivative of cost with respect to equity capital is greater than or equal to -0.18, rejection of which would suggest over-utilization of equity capital, and the null that the derivative is less than or equal to -0.14, rejection of which would suggest that equity capital is under-utilized. We performed these tests at levels 0.01 and 0.1; results are displayed in Table D.1.

For the vast majority of institutions in 2006.Q4, we find evidence of over-utilization of equity capital. Our results are similar to findings by Hughes et al. (2001) for banks up to about $10 billion of assets, but we find that larger banks also tend to over-utilize equity capital. To the extent that our results may differ from those of Hughes et al., this may be due to the fact that Hughes et al. specify and estimate a translog functional form for

costs, which we have shown to severely mis-specify the cost relation. In addition, Hughes et al. suggest that most banks with assets ranging from $10 billion to $50 billion use optimal levels of capital, but they arrive at this conclusion since neither null hypothesis is rejected for most banks in this size range. However, it is important to remember that failure to reject a null hypothesis can happen for many reasons, and does not by itself imply that the null is true. In any case, our results suggest that banks are not allocatively efficient to the extent that they employ too much equity capital.

References

Aitchison, J. and C. G. G. Aitken (1976), Multivariate binary discrimination by the kernel method, *Biometrika* 63, 413–420.

Bierens, H. J. (1987), Kernel estimators of regression functions, in T. F. Bewley, ed., *Advances in Econometrics, Fifth World Congress*, volume 1, Cambridge: Cambridge University Press, pp. 99–144.

Delgado, M. A. and J. Mora (1995), On asymptotic inferences in nonparametric and semiparametric models with discrete and mixed regressors, *Investigaciones Economicas* 19, 435–467.

Efron, B. and R. J. Tibshirani (1993), *An Introduction to the Bootstrap*, London: Chapman and Hall.

Fan, J. and I. Gijbels (1994), Censored regression: Local linear regression smoothers, *Journal of the American Statistical Association* 89, 560–570.

— (1996), *Local Polynomial Modelling and Its Applications*, London: Chapman and Hall.

Fan, J., M. B. T. Gasser, I. Gijbels, and J. Engel (1997), Local polynomial regression: Optimal kernels and asymptotic minimax efficiency, *Annals of the Institute for Statistical Mathematics* 49, 79–99.

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

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

Hughes, J. P., L. J. Mester, and C. G. Moon (2001), Are scale economies in banking elusive or illusive? evidence obtained by incorporating capital structure and risk-taking into models of bank production, *Journal of Banking and Finance* 25, 2169–2208.

Loftsgaarden, D. O. and C. P. Quesenberry (1965), A nonparametric estimate of a multivariate density function, *Annals of Mathematical Statistics* 36, 1049–1051.

Mammen, E. (1992), *When Does Bootstrap Work? Asymptotic Results and Simulations*, Berlin: Springer-Verlag.

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

Pagan, A. and A. Ullah (1999), *Nonparametric Econometrics*, Cambridge: Cambridge University Press.

Racine, J. and Q. Li (2004), Nonparametric estimation of regression functions with both categorical and continuous data, *Journal of Econometrics* 119, 99–130.

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

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

White, H. (1980), A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity, *Econometrica* 48, 817–838.

Wilson, P. W. and K. Carey (2004), Nonparametric analysis of returns to scale and product mix among US hospitals, *Journal of Applied Econometrics* 19, 505–524.

Table C.3: Expansion-Path Scale Economies (99-Percent Significance)

Year	IRS	CRS	DRS
1984	12422	6	0
1995	9278	4	0
2006	6893	68	0

Table C.4: Summary Statistics for Expansion-Path Scale Economy Estimates by Size-Quartile (99-Percent Significance

Size Quartile	Min	1st Quartile	Median	Mean	3rd Quartile	Max
			—1984.Q4—			
1	0.8730	0.9365	0.9420	0.9416	0.9471	0.9723
2	0.9115	0.9401	0.9449	0.9449	0.9496	0.9780
3	0.9123	0.9396	0.9444	0.9445	0.9494	0.9851
4	0.9118	0.9407	0.9451	0.9452	0.9495	0.9772
			—1995.Q4—			
1	0.9078	0.9366	0.9425	0.9423	0.9482	0.9734
2	0.9141	0.9395	0.9450	0.9450	0.9508	0.9772
3	0.9057	0.9396	0.9450	0.9451	0.9504	0.9823
4	0.8887	0.9394	0.9447	0.9447	0.9497	0.9818
			—2006.Q4—			
1	0.8788	0.9330	0.9400	0.9393	0.9460	0.9997
2	0.9026	0.9371	0.9428	0.9427	0.9488	0.9883
3	0.9094	0.9394	0.9452	0.9456	0.9516	0.9847
4	0.9064	0.9397	0.9461	0.9463	0.9524	1.0005

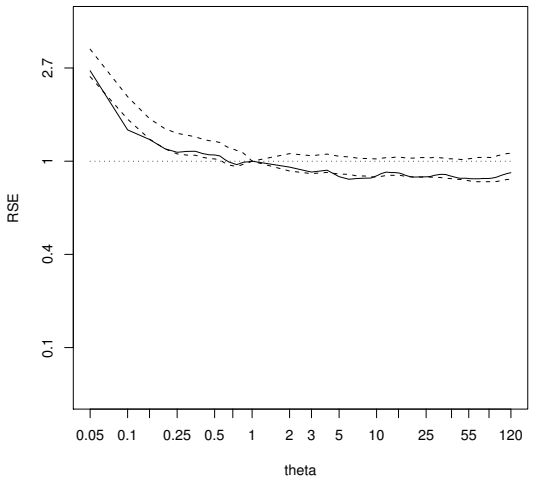
NOTE: For each period, summary statistics are given for the first, second, third, and fourth quartiles of banks' total assets.

Table D.1: Test of First-Order Conditions for Cost-Minimizing Level of Equity Capital (Sample #2, 2006.Q4)

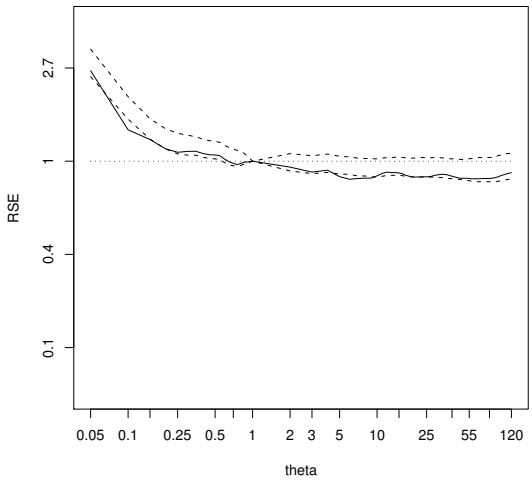
	1%			10%		
	Under	No Reject	Over	Under	No Reject	Over
full sample	0.5	1.1	98.4	0.6	0.5	98.8
≤$300 million	0.4	0.7	98.9	0.4	0.3	99.2
$300 million — $2 billion	0.9	2.7	96.4	1.4	1.3	97.2
$2 billion – $10 billion	1.6	3.2	95.2	1.6	1.6	96.8
$10 billion – $50 billion	2.5	2.5	95.0	2.5	2.5	95.0
>$50 billion	0.0	5.3	94.7	5.3	0.0	94.7

NOTE: The values in this table give the percentages of observations in sample #2 for which inference at either 1 or 10 percent levels indicates that equity capital is under- or over-utilized for an (unobserved) price of equity capital in the range $[0.14, 0.18]$. The "under" columns report the percentages of observations for which the null hypothesis $\partial\mathcal{C}/\partial Z2 + 0.18 \geq 0$ is rejected. The "over" columns report the percentages of observations for which the null hypothesis $\partial\mathcal{C}/\partial Z2 + 0.14 \leq 0$ is rejected. The columns labeled "no reject" report the percentages of observations for which neither of these null hypotheses can be rejected; of course, failure to reject a null hypothesis does not imply that the null is true; a statistical test can only either reject or fail to reject the null. Sample #2 contains 6,090 observations for 2006.Q4.

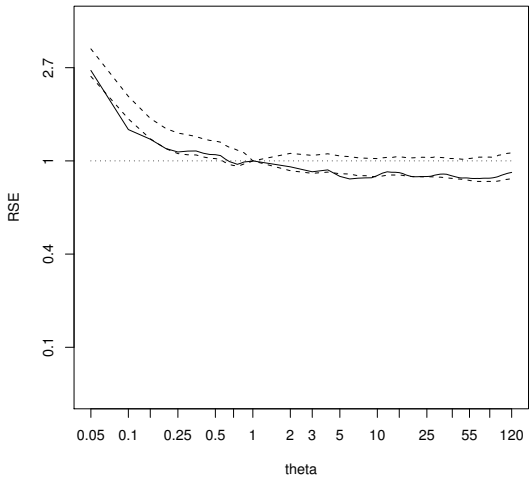
Figure C.2: Ray Scale Economies (MBHC = 1, year 1984



Statewide Branching

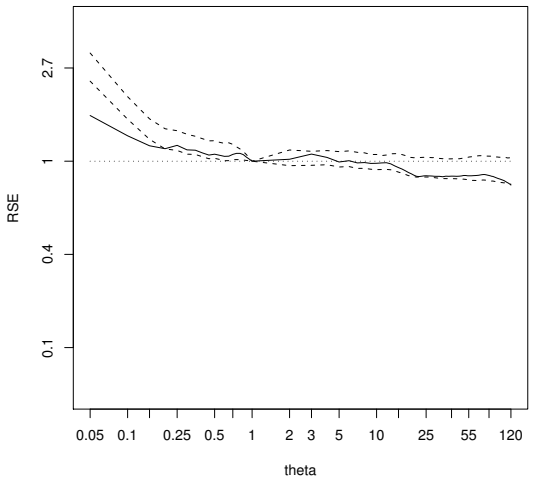


Limited Branching

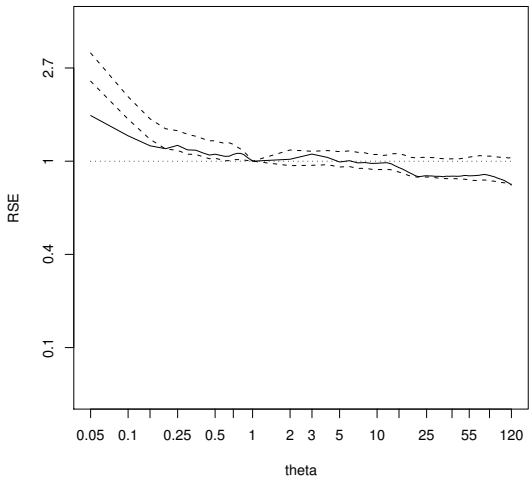


Unit Banking

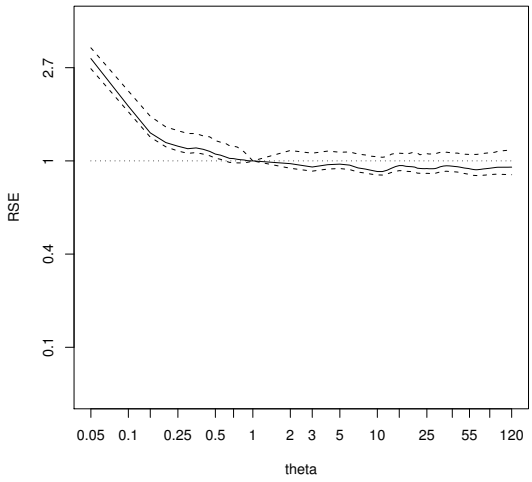
Figure C.3: Ray Scale Economies (MBHC = 1, years 1995 and 2006



Statewide Branching, 1995



Limited Branching, 1995



Interstate Branching, 2006

Figure C.4: Expansion Path Scale Economies by Size-Quartile, 1984

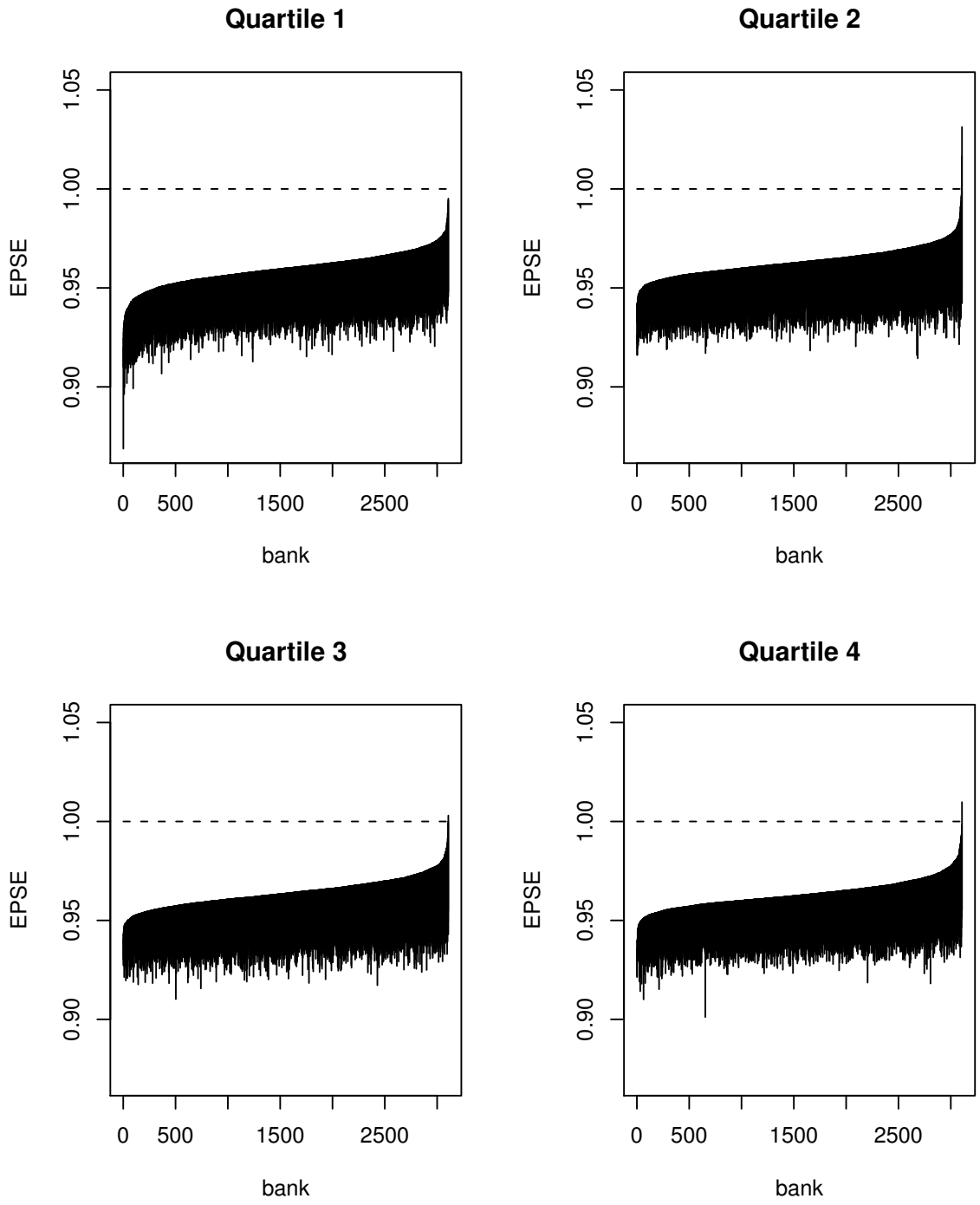


Figure C.5: Expansion Path Scale Economies by Size-Quartile, 1995

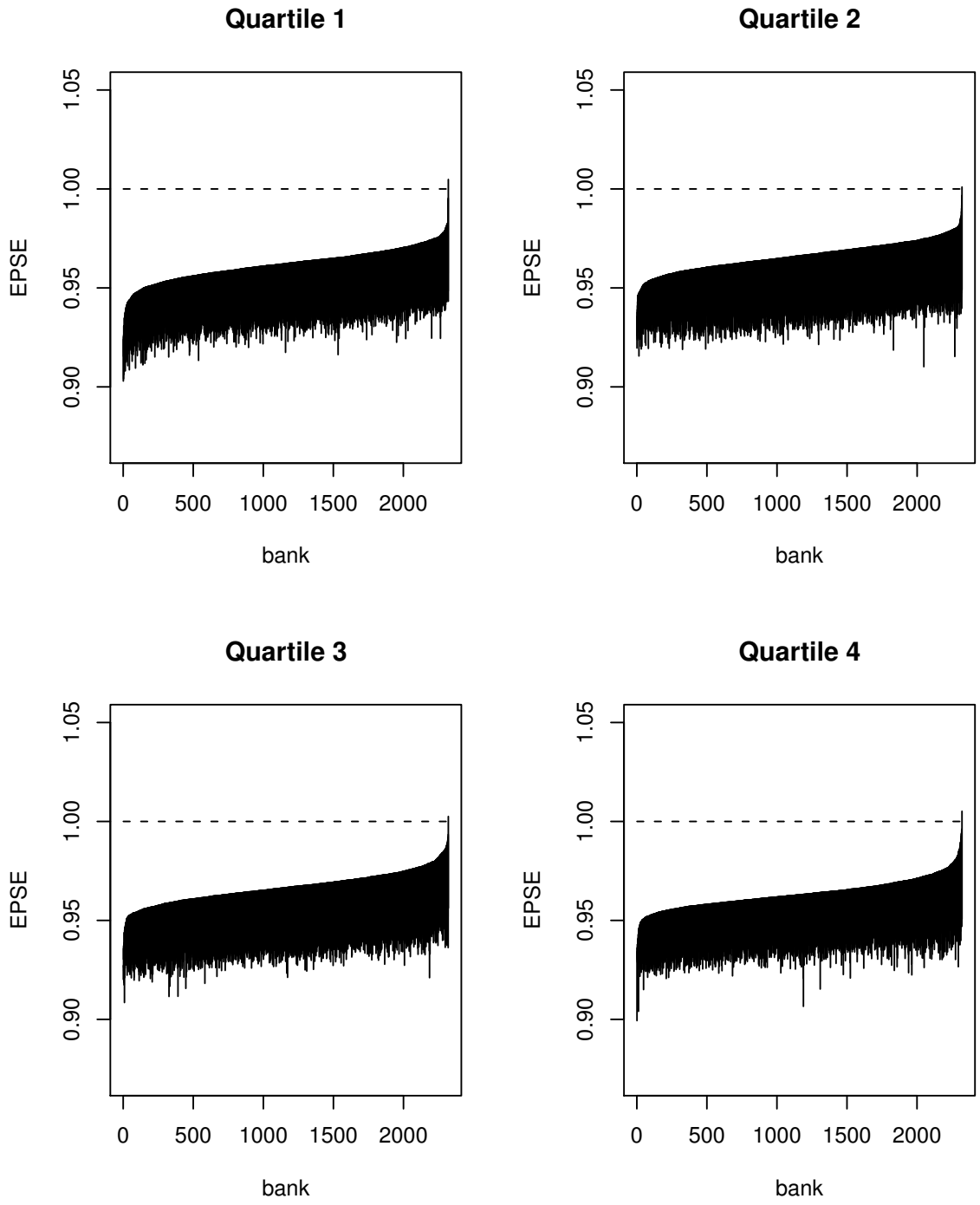


Figure C.6: Expansion Path Scale Economies by Size-Quartile, 2006