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

\[
\log(COST/W3_i) = \beta_0 + \sum_{j=1}^{9} \beta_j X_{ij} + \sum_{j=1}^{9} \sum_{k=1}^{j} \beta_{jk} X_{ij} X_{ik} + \varepsilon
\]  

(A.1)

where \(E(\varepsilon) = 0\) and \(X_i\) contains the \(i\)th 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 \(\beta_j = [\beta_1 \ldots \beta_{55}]^t\), and let \(X_j\) be the \((n_j \times 55)\) matrix containing the right-hand side variables in (A.1); the first column of \(X_j\) consists of a vector of 1s. In addition, let \(Y_j\) 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

\[
Y_j = X_j \beta_j + \varepsilon_j, \quad \text{(A.2)}
\]

where \(\varepsilon_j\) 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 \( \hat{\beta}_j \) and \( \hat{\varepsilon}_j = Y_j - X_j \hat{\beta}_j \). Next, we compute White’s (1980) heteroskedasticity-consistent covariance matrix estimator

\[
\hat{\Sigma}_j = (X'_jX_j)^{-1} (X'_j \text{diag}(\hat{\varepsilon}^2_j) X_j) (X'_j X_j)^{-1}
\]  

for each sub-sample, where \( \text{diag}(\hat{\varepsilon}^2_j) \) is the \((n_j \times n_j)\) diagonal matrix with elements of \( \hat{\varepsilon}^2_j \) along the principal diagonal. Under the null hypothesis \( H_0: \beta_1 = \beta_2 \), asymptotic normality of OLS estimators ensures that the Wald statistic

\[
\hat{W} = (\hat{\beta}_1 - \hat{\beta}_2)' (\hat{\Sigma}_1 + \hat{\Sigma}_2)^{-1} (\hat{\beta}_1 - \hat{\beta}_2) \xrightarrow{d} \chi^2(55).
\]  

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 \( 3 \times 10^{-163} \). The largest p-value allows us to reject the translog specification at 5 percent significance; the next-largest p-value was \( 2.488 \times 10^{-5} \). 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 \( V \) define the function

\[
\psi_1(V) \equiv (V - n^{-1}i'iV) (n^{-1}V'iV - n^{-2}V'i'i'iV)^{-1/2}
\]  

For an \((n \times 1)\) vector \( V \) define the function
where \( i \) 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 \( A \) be an \((n \times 10)\) matrix with columns \( \psi_1(\log(1 + Y_1)), \psi_1(\log(1 + Y_2)), \psi_1(\log(1 + Y_3)), \psi_1(\log(Y_4)), \psi_1(\log(Z_1)), \psi_1(\log(1 + Z_2)), \psi_1(\log(W_1/W_3)), \psi_1(\log(W_2/W_3)), \psi_1(\log(5 + \text{OFF})), \) and \( \psi_1(\log(\text{TIME})) \).

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

\[
\phi_k = \frac{\sum_{j=1}^{k} \lambda_j}{\sum_{\ell=1}^{10} \lambda_\ell}
\]  

(B.2)

represents the proportion of the independent linear information in \( A \) that is contained in the first \( k \) principal components, i.e., the columns of \( P \).

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 \( P_k \) denote the \( k \)th column of the principal component matrix \( P \) and define

\[
\psi_0(P_k) \equiv P_k \left[ n^{-1} P_k' P_k - n^{-2} P_k' i i' P_k \right]^{-1/2}.
\]  

(B.3)

The transformation \( \psi_0(P_k) \) has (constant) unit variance. Next, let \( z_i \) represent the row vector containing the \( i \)th observations on \( \psi_0(P_{1}), \ldots, \psi_0(P_{6}) \). In addition, let \( u_i \) represent the row vector containing the \( i \)th observations on the binary variables MBHC, STATEWIDE, LIMITED, and UNIT. We can now write our model as the following regression equation:

\[
C_i = m(z_i, u_i) + \xi_i
\]  

(B.4)
where the subscript \( i \) indexes observations, \( C_i = \psi_1 (\log(\text{COST}/W3)) \), \( \xi_i \) is a random error term with \( E(\xi_i) = 0 \), and \( \text{VAR}(\xi_i) = \sigma^2(z_i, u_i) \). The function \( m(z_i, u_i) = E(C_i | z_i, u_i) \) is a conditional mean function and can be estimated by non-parametric methods. Moreover, since the transformation from COST to \( C \) can be inverted, given an estimated value \( \hat{m}(z, u) \), straightforward algebra leads to an estimate
\[
\hat{C}(y, w) = \exp \left[ \psi_1^{-1}(\hat{m}(z, u)) \right].
\] (B.5)

As discussed below, we use a local linear estimator to estimate \( m(z, u) \). Although this estimator is weakly consistent, it is asymptotically biased. Moreover, even if \( \hat{m}(z, u) \) were unbiased, use of the nonlinear transformation in (B.5) means that \( \hat{C}(y, w) \) 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 \( C(y, w) \) were used, plugging the estimator into the definitions of \( S \) and \( E_0 \) given in the text to obtain estimators \( \hat{S} \) and \( \hat{E}_0 \) involves additional nonlinear transformations. Fortunately, any bias in the resulting estimates \( \hat{S} \) and \( \hat{E}_0 \) 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 \( m(z) \). 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 \( m_*(z) \). Note that \( z \) is a vector of length \( \ell \). The local linear estimator follows from a first-order Taylor expansion of \( m_*(z) \) in a neighborhood
of an arbitrary point \( z_0 \):

\[
m_*(z) \approx m_*(z_0) + \frac{\partial m_*(z_0)}{\partial z}(z - z_0).
\]

This suggests estimating the conditional mean function at \( z_0 \) by solving the locally weighted least squares regression problem

\[
[\hat{\alpha}_0 \hat{\alpha}]' = \arg\min_{\alpha_0, \alpha} \sum_{i=1}^{n} [C_i - \alpha_0 - (z_i - z_0)\alpha]^2 K (|H|^{-1}(z_i - z_0))
\]

where \( K(\cdot) \) is a piece-wise continuous multivariate kernel function satisfying \( \int_{\mathbb{R}^\ell} K(u) du = 1 \) and \( K(u) = K(-u) \); \( H \) is an \( \ell \times \ell \) matrix of bandwidths; \( \alpha_0 \) is a scalar, and \( \alpha \) is an \( \ell \)-vector.

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

\[
[\hat{\alpha}_0 \hat{\alpha}]' = (Z'\Phi Z)^{-1} Z'\Phi C,
\]

where \( C = [C_1 \ldots C_n]' \), \( \Phi = \text{diag}[K(|H|^{-1}(z_i - z_0))] \), and \( Z \) is an \( n \times (\ell + 1) \) matrix with \( i \)th row given by \( [1 \ (z_i - z_0)] \). The fitted value \( \hat{\alpha}_0 \) provides an estimate \( \hat{m}_*(z_0) \) of the conditional mean function \( m_*(z_0) \) at an arbitrary point \( z_0 \).

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 \( \Phi \) 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)

\[\text{The fitted values in } \hat{\alpha} \text{ provide estimates of elements of the vector } \frac{\partial m(z_0)}{\partial z}. \text{ However, if the object is to estimate first derivatives, mean-square error of the estimates can be reduced by locally fitting a quadratic rather than a linear expression (see Fan and Gijbels, 1996 for discussion); this increases computational costs, which are already substantial for the local linear fit. Moreover, determining the optimal bandwidths becomes more difficult and computationally more burdensome for estimation of derivatives. See Härdle (1990, pp. 160–162) for discussion of some of the issues that are involved with bandwidth selection for derivative estimation.}\]
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 \( u_i \) represent a vector of observations on \( k = 4 \) binary dummy variables, and consider an arbitrary Bernoulli vector \( u_0 \) of length \( k \). Then let \( \delta(u_i, u_0) = (u_i - u_0)'(u_i - u_0) \), and define the discrete kernel function

\[
G(u_i \mid u_0, h_1) = h_1^{k-\delta(u_i, u_0)} (1 - h_1)^{\delta(u_i, u_0)}
\]

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

Note that \( \lim_{h_1 \to 1} G(u_0 \mid u_i, h_1) \) equals either 1 or 0, depending on whether \( u_0 = u_i \) or \( u_0 \neq u_i \), 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 \( G(u_0 \mid u_i, h_1) = 1 \) regardless of whether \( u_0 = u_i \) or \( u_0 \neq u_i \); 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 \( \ell \)-variate spherically symmetric Epanechnikov kernel with a single, scalar bandwidth \( h_0 \); i.e.,

\[
K(t) = \frac{\ell(\ell + 2)}{2S_\ell} (1 - t't)I(t't \leq 1)
\]

where \( I(\cdot) \) is the indicator function, \( S_\ell = 2\pi^{\ell/2}/\Gamma(\ell/2) \), \( \Gamma(\cdot) \) denotes the gamma function, \( \mathbf{u} = |\mathbf{H}|^{-\ell}(\mathbf{z}_i - \mathbf{z}_0) \), and \( \mathbf{H} \) 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 \( \mathcal{D} = \{0, 1\} \times \{(1, 0, 0), (0, 1, 0), (0, 0, 1), (0, 0, 0)\} \) be the set of possible values for the vector \( \mathbf{u} \) of binary variables. Incorporating the discrete covariates, an estimate \( \hat{m}(\mathbf{z}_0, \mathbf{u}_0) \) of
the conditional mean function in (B.4) at an arbitrary point \((z_0, u_0) \in \mathbb{R}^\ell \times \mathbb{D}\) is given by 
\[
\hat{\alpha}_0 \text{ obtained from }
\begin{bmatrix}
\hat{\alpha}_0 \\
\hat{\alpha}
\end{bmatrix}' = \arg\min_{\alpha_0, \alpha} \sum_{i=1}^n [C_i - \alpha_0 - (z_i - z_0)\alpha]^2 K(\ell^{-1}(z_i - z_0)) G(u_i | u_0, h_1) \tag{B.11}
\]
where \(u_0 \in \mathbb{D}\). The solution to the least-squares problem in (B.11) is given by
\[
\begin{bmatrix}
\hat{\alpha}_0 \\
\hat{\alpha}
\end{bmatrix}' = (Z'\Omega Z)^{-1} Z'\Omega C, \tag{B.12}
\]
where \(Z\) is defined as in (B.8) and the matrix \(\Omega\) of weights is given by
\[
\Omega = \text{diag}[K(\ell^{-\ell}(z_i - z_0)) G(u_i | u_0, h_1)]. \tag{B.13}
\]

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 \(Z\) have constant, unit variance, suggesting use of the same bandwidth in each direction. Hence we set \(H = \text{diag}(h(z_0))\) so that \(|H|^{-\ell} = h(z_0)^{-\ell}\), where \(h(z_0)\) is an adaptive scalar bandwidth depending on the point \(z_0\) 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 \(h(z_0)\) 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 \(h(z_0)\) for any particular point \(z_0 \in \mathbb{R}^\ell\) as the maximum Euclidean distance between \(z_0\) and the \(\kappa\) nearest points in the observed sample \(\{z_i\}^{n}_{i=1}\), \(\kappa \in \{2, 3, 4, \ldots\}\). Thus, given the data and the point \(z_0\), the bandwidth \(h(z_0)\) is determined by \(\kappa\), and varies depending on the density of the continuous explanatory variables locally around the point \(z_0 \in \mathbb{R}^\ell\) 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, \(z_0\). 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 \(\kappa\) 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

$$\left(\hat{\omega}, \hat{h}_1\right) = \arg\min_{\omega, h_1} \sum_{i=1}^{n} \left[C_i - \hat{m}_{-i}(z_i, u_i)\right]^2,$$

where $\hat{m}_{-i}(z_i, u_i)$ is computed the same way as $\hat{m}(z_i, u_i)$, except that the $i$th diagonal element of $\Psi$ 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 $(z_0, u_0) \in \mathbb{R}^\ell \times \mathbb{D}$. We then estimate the RSE and EPSE measures defined 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 $\kappa$ 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.
which we substitute into the definitions of $S$ and $E_0$ in the text to obtain bootstrap values $\hat{S}_b^*$ and $\hat{E}_0^*$ for particular values of $z$ and $u$, with $b = 1, \ldots, B$.

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

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

$$
\alpha_2 = \Phi\left(\hat{\varphi}_0 + \frac{\varphi(1-\alpha/2)}{1 - \hat{\varphi}_0 + \varphi(1-\alpha/2)}\right),
$$
(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

$$
\hat{\varphi}_0 = \Phi^{-1}\left(\frac{\#\{\hat{S}_b^* < \hat{S}\}}{B}\right),
$$
(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 $\{(\hat{S}^*_b - \hat{S})\}_{b=1}^B$ by algebraic value, delete $(\alpha/2 \times 100)$-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 $S$ is

$$
\hat{S} + a^*_\alpha \leq S \leq \hat{S} + b^*_\alpha.
$$
(B.18)

The idea underlying (B.18) is that the empirical distribution of the bootstrap values $\left(\hat{S}_b^* - \hat{S}\right)$ mimics the unknown distribution of $\left(\hat{S} - S\right)$, 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 $\left(\hat{S}_b^* - \hat{S}\right)$ (we set $B = 1000$). Any bias in $\hat{S}$ relative to $S$ is reflected in bias of $\hat{S}^*$ relative to $\hat{S}$; in the case of large bias, it is conceivable that the estimated confidence interval may not include the original estimate $\hat{S}$, since the estimated confidence interval corrects for the bias in $\hat{S}$. We estimate confidence intervals for the EPSE measures similarly.

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 policymakers. 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
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


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

<table>
<thead>
<tr>
<th>Year</th>
<th>IRS</th>
<th>CRS</th>
<th>DRS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1984</td>
<td>12422</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>1995</td>
<td>9278</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>2006</td>
<td>6893</td>
<td>68</td>
<td>0</td>
</tr>
</tbody>
</table>
Table C.4: Summary Statistics for Expansion-Path Scale Economy Estimates by Size-Quartile (99-Percent Significance)

<table>
<thead>
<tr>
<th>Size Quartile</th>
<th>1st Quartile</th>
<th>Median</th>
<th>Mean</th>
<th>3rd Quartile</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>1984.Q4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.8730</td>
<td>0.9420</td>
<td>0.9416</td>
<td>0.9471</td>
<td>0.9723</td>
</tr>
<tr>
<td>2</td>
<td>0.9115</td>
<td>0.9449</td>
<td>0.9449</td>
<td>0.9496</td>
<td>0.9780</td>
</tr>
<tr>
<td>3</td>
<td>0.9123</td>
<td>0.9444</td>
<td>0.9445</td>
<td>0.9494</td>
<td>0.9851</td>
</tr>
<tr>
<td>4</td>
<td>0.9118</td>
<td>0.9451</td>
<td>0.9452</td>
<td>0.9495</td>
<td>0.9772</td>
</tr>
<tr>
<td>1995.Q4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.9078</td>
<td>0.9425</td>
<td>0.9423</td>
<td>0.9482</td>
<td>0.9734</td>
</tr>
<tr>
<td>2</td>
<td>0.9141</td>
<td>0.9450</td>
<td>0.9450</td>
<td>0.9508</td>
<td>0.9772</td>
</tr>
<tr>
<td>3</td>
<td>0.9057</td>
<td>0.9451</td>
<td>0.9451</td>
<td>0.9504</td>
<td>0.9823</td>
</tr>
<tr>
<td>4</td>
<td>0.8887</td>
<td>0.9447</td>
<td>0.9447</td>
<td>0.9497</td>
<td>0.9818</td>
</tr>
<tr>
<td>2006.Q4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.8788</td>
<td>0.9400</td>
<td>0.9393</td>
<td>0.9460</td>
<td>0.9997</td>
</tr>
<tr>
<td>2</td>
<td>0.9026</td>
<td>0.9428</td>
<td>0.9427</td>
<td>0.9488</td>
<td>0.9883</td>
</tr>
<tr>
<td>3</td>
<td>0.9094</td>
<td>0.9456</td>
<td>0.9456</td>
<td>0.9516</td>
<td>0.9847</td>
</tr>
<tr>
<td>4</td>
<td>0.9064</td>
<td>0.9461</td>
<td>0.9463</td>
<td>0.9524</td>
<td>1.0005</td>
</tr>
</tbody>
</table>

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)

<table>
<thead>
<tr>
<th></th>
<th>1% No Reject</th>
<th>10% No Reject</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Under Reject</td>
<td>Over</td>
</tr>
<tr>
<td>full sample</td>
<td>0.5</td>
<td>1.1</td>
</tr>
<tr>
<td>≤$300 million</td>
<td>0.4</td>
<td>0.7</td>
</tr>
<tr>
<td>$300 million — $2 billion</td>
<td>0.9</td>
<td>2.7</td>
</tr>
<tr>
<td>$2 billion – $10 billion</td>
<td>1.6</td>
<td>3.2</td>
</tr>
<tr>
<td>$10 billion – $50 billion</td>
<td>2.5</td>
<td>2.5</td>
</tr>
<tr>
<td>&gt;$50 billion</td>
<td>0.0</td>
<td>5.3</td>
</tr>
</tbody>
</table>

**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 $\frac{\partial C}{\partial Z2} + 0.18 \geq 0$ is rejected. The “over” columns report the percentages of observations for which the null hypothesis $\frac{\partial C}{\partial Z2} + 0.14 \leq 0$ 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

Quartile 1

Quartile 2

Quartile 3

Quartile 4
Figure C.5: Expansion Path Scale Economies by Size-Quartile, 1995
Figure C.6: Expansion Path Scale Economies by Size-Quartile, 2006

Quartile 1

Quartile 2

Quartile 3

Quartile 4

bank

bank

bank

bank