Working Paper 9117

INEFFICIENCY AND PRODUCTIVITY GROWTH IN BANKING: A COMPARISON OF STOCHASTIC ECONOMETRIC AND THICK FRONTIER METHODS

by Paul W. Bauer, Allen N. Berger,
and David B. Humphrey

Paul W. Bauer is an economist at the Federal Reserve Bank of Cleveland, Allen N. Berger is a senior economist at the Board of Governors of the Federal Reserve System, and David B. Humphrey is a professor of finance at Florida State University. The authors would like to thank Knox Lovell and James Thomson for helpful comments, and Alex Wolman and Fadi Alameddine for outstanding research assistance.

December 1991

I. Introduction

Until recently, bank cost studies focused almost exclusively on scale and product mix (scope) economies. While this approach has been useful, a potentially more important dimension of bank cost economies appears to be differences in efficiency. Recent studies have estimated inefficiencies of 20 percent or more of costs, even for banks of similar scale and product mix. These inefficiencies appear to dominate scale and product mix effects, which usually average 5 percent or less.

Our purposes in this paper are twofold. First, we apply two methods of efficiency measurement that have been employed in extant efficiency literature and contrast the results across methodologies. Specifically, we compare the stochastic econometric frontier approach, first proposed by Aigner, Lovell, and Schmidt (1977), Meeusen and van den Broeck (1977), and Battese and Corra (1977), with the thick frontier approach of Berger and Humphrey (1991a,b). These methods are applied to a panel data set of 683 U.S. branching state banks with over $100 million in assets that were continuously in existence during 1977-88. Since these institutions account for two-thirds of total U.S. banking assets, and all U.S. states allow some form of branching as of August 1991, our results may be considered reasonably representative of the industry as a whole. Since we have data for each year, we are also able to assess the variations in efficiency over time. Previous banking studies have not compared the results of alternative frontier methods (Ferrier and Lovell [1990] excepted), nor have they assessed efficiency at more than one point in time (Berger and Humphrey [1991b] excepted).

The second purpose of the paper is the measurement of the growth of total factor productivity (TFP) in banking, which incorporates both technical change and scale economies. While some prior studies have investigated technical change in banking, they have done so using data for all banks, rather than for those on the efficient frontier. Such a procedure may confound technical change on the frontier with fluctuations in inefficiency that alter the average distance from the frontier. These prior studies also have determined *average* time trends for technical change, rather than specific year-to-year variations. This

study breaks both of these molds by estimating annual shifts for the efficient frontier, rather than for the universe of banks, and by permitting the size of these shifts to vary freely on a year-to-year basis. TFP growth is determined by combining our frontier measures of technical change with scale economy measures for frontier banks. In this way, productivity growth, or movement *of* the frontier, is considered separately from changes in inefficiency, or average distance *from* the frontier. Productivity growth over 1977-88 is of particular interest because of financial market innovation (cash management), regulatory change (deregulation of consumer deposit rates), and technical innovation (automated teller machines) during this interval.

II. A Brief Comparison of Frontier Efficiency Models

Inefficiency is assessed by measuring how far a firm's costs or input requirements deviate from a "best practice" set of firms, or an efficient frontier. The key methodological problem is that the true technically based frontier is unknown and must be estimated from levels found in the data set. The differences among techniques in the efficiency literature largely reflect differing maintained assumptions used in estimating these frontiers.

The stochastic econometric frontier approach modifies a standard cost (or production) function to allow inefficiencies to be included in the error term. A composite error term is specified that includes both random error and inefficiency, and specific distributional assumptions are made to separate these two components. Since inefficiencies only increase costs above frontier levels, while random fluctuations can either increase or decrease costs, inefficiencies are assumed to be drawn from a one-sided distribution (usually the half-normal), and random fluctuations are assumed to be drawn from a symmetric distribution (usually the normal). This approach has been applied to banking by Ferrier and Lovell (1990). Berger (1991) used different techniques, described below, to identify the inefficiencies.

The thick frontier approach, instead of estimating a frontier *edge*, compares the average efficiencies of large groups of banks. Banks in the lowest average cost quartile are assumed to have above-average efficiency and to form a thick frontier. Similarly, banks in the highest average cost quartile are identified as likely having below-average efficiency. Differences in

error terms within the highest and lowest cost quartiles are assumed to reflect random error, while the predicted cost differences between these quartiles are assumed to reflect inefficiencies plus exogenous differences in output quantities and input prices. Banks are stratified by size class before the quartiles are formed both to ensure that a broad range of institutions are represented in each quartile and to reduce the relationship between the quartile selection criterion and the dependent variable in the regressions.[1] The thick frontier approach has been applied to banking by Berger and Humphrey (1991a,b).

The data envelopment analysis (DEA) approach, which is not replicated here, assumes that random error is zero so that *all* unexplained variations are treated as reflecting inefficiencies. The DEA approach has been applied to banking by Rangan, Grabowski, Aly, and Pasurka (1988), Aly, Grabowski, Pasurka, and Rangan (1990), Elyasiani and Mehdian (1990), and Ferrier and Lovell (1990).

An illustrative comparison of the stochastic and thick frontier methods can be made using the raw data presented in Figure 1A, which shows how average total costs per dollar of assets varies across eight bank size classes. The AC_{Q1} and AC_{Q4} lines are average costs for the lowest and highest cost quartiles respectively, while the AC_{MIN}, AC_{MAX}, and AC_{MEAN} lines correspond to the overall minima, maxima, and means respectively. The averages for all of these curves are taken over the 12-year period from 1977-88, and so do not reflect the full variation in costs for any one year.[2] Mean average costs (AC_{MEAN}) are very flat across different sized banks--the range of variation is only 5 percent--suggesting few scale economies or diseconomies. The average costs of the lowest and highest quartiles (AC_{Q1} and AC_{Q4}) are also relatively flat, with ranges of variation of 5 and 13 percent respectively.

---

1. The quartile selection criterion could bias the coefficient estimates if the dependent variable fluctuates too closely with the criterion variable (average costs by size class). This does not appear to be a problem here, since a regression of our dependent variable, the log of total costs, on dummy variables for the cost quartiles yielded an R^2 of less than .01. The R^2 is low because size is the main determinant of the dependent variable, and most of the effects of size are removed when the quartiles are formed separately by size class.

2. For example, AC_{MIN} represents the costs of the bank with the lowest long-run costs in each size class, but in any one year, some other banks had short-run costs one-third to one-half as high.

In the thick frontier approach, cost equations are estimated for the highest and lowest quartiles, and the difference in predicted costs for a given set of output quantities and input prices is considered to be due to inefficiency. The raw-data approximation to this inefficiency is given in Figure 1A by the difference [AC_{Q4} - AC_{Q1}], which averages 23 percent. Measured inefficiency will differ from this because the cost equations control for and net out differences in output scale, product mix, and input prices.

The stochastic econometric frontier will lie somewhere between the minimum costs and the mean of the data, or approximately between AC_{MIN} and AC_{MEAN}. The precise location depends upon the actual shape of the distribution of the data and the assumed distribution for the inefficiencies. If the data are skewed toward the higher cost banks, the stochastic econometric frontier will tend toward AC_{MIN}, while if the data are relatively unskewed, the frontier will lie closer to AC_{MEAN}. A raw-data estimate of the maximum average inefficiency under this method is given in Figure 1A by the difference [AC_{MEAN} - AC_{MIN}], which averages 29 percent. Note that the inefficiencies between the two approaches given here are not strictly comparable, since the thick frontier approach compares high cost and low cost banks, while the econometric frontier approach takes the average distance of all banks from the frontier edge. However, these methods will be made comparable below.

Figure 1B presents a time series of average costs by bank quartile over time. Since the costs in the numerator and the assets in the denominator are both in nominal terms, the effects of inflation net out, and the curves reflect the time trend of real average cost. The inverse of real average cost, corrected for changes in input prices, indicates the trend in bank technical change. Real average cost for Q1 banks (AC_{Q1}) rose 34 percent over 1977-88, while the growth in input prices was generally less than this amount, suggesting that technical change may have been negative or very low over this period. This conclusion is unlikely to be altered by considering the effect of scale economies on TFP measurement, since these economies appear to be small. As discussed below, negative or low TFP growth may be associated with the peculiar nature of the banking industry's response to payments developments of the late 1970s and deposit rate deregulation and technical innovations of the 1980s.

The formal cost model underlying both the stochastic econometric and thick frontier approaches used here can be written as:

$$\begin{aligned} \ln TC = {} & \alpha + \sum_{i=1}^{5} \beta_i \ln Y_i + 1/2 \sum_{i=1}^{5}\sum_{j=1}^{5} \delta_{ij} \ln Y_i \ln Y_j + \sum_{k=1}^{4} \gamma_k \ln P_k + \\ & 1/2 \sum_{k=1}^{4}\sum_{n=1}^{4} \gamma_{kn} \ln P_k \ln P_n + \sum_{i=1}^{5}\sum_{k=1}^{4} \rho_{ik} \ln Y_i \ln P_k + \lambda_M M + \lambda_U U + \\ & \sum_{i=1}^{5} \theta_{iM} \ln Y_i M + \sum_{i=1}^{5} \theta_{iU} \ln Y_i U + \sum_{k=1}^{4} \phi_{kM} \ln P_k M + \varepsilon \end{aligned} \quad (1)$$

$$S_k = \alpha_k + \sum_{n=1}^{4} \gamma_{kn} \ln P_n + \sum_{i=1}^{5} \rho_{ik} \ln Y_i + \phi_{kM} M + \psi_k, \quad k=1,2,3,4 \quad (2)$$

where:

TC = real total cost (interest and operating costs deflated by the GNP deflator);[3]

Y_i = real value of output i: 1) demand deposits, 2) small time and savings deposits, 3) real estate loans, 4) commercial and industrial loans, and 5) installment loans;

P_k = real price of input k: 1) labor, 2) physical capital, 3) interest rate on small time and savings deposits, and 4) interest rate on purchased funds;

M = bank merger dummy variable, equals 1 for a bank in the year of its merger;[4]

U = unit banking dummy variable, equals 1 when the bank was subject to unit laws (we include data from 8 states that adopted branching laws during 1977-88);

S_k = cost share of input k, which equals $\partial \ln TC/\partial \ln P_k$ from (1) plus an error term;

ε, ψ_k = error terms.[5]

3. As is standard in banking studies, cost figures do not include loan losses. They are instead effectively treated as declines in revenue, since the rates charged on loans include premia to cover the expected value of these losses.

4. While the effects of a merger may last beyond the year it occurs, the results were materially unchanged when this dummy variable was respecified to be 1 in the year of the merger and in all following years.

5. The standard symmetry and linear homogeneity in input prices restrictions are imposed in estimation, as are the Shephard's Lemma cross-equation restrictions. One of the share equations in (2) is dropped to avoid singularity. The number of branch offices is not specified in this model because cost efficiencies at the level of the banking firm are desired, rather than those for the average office. The relatively unimportant interactions between the U term and the $\ln P_k$ and M terms are not specified in order to conserve on the number of parameters to be estimated.

The key to the stochastic econometric approach is the two-part composed error term, one part for inefficiency and one for statistical noise. Thus, $\varepsilon = \mu+v$, where $\mu \geq 0$ represents inefficiency and v represents independent statistical noise. In our panel estimations, v is allowed to follow a first-order serial correlation process. Unless panel data are available, one must assume that the regressors are independent of the inefficiency term as well as of the statistical noise. In addition, specific distributions for the noise and inefficiency terms must be imposed. However, if panel data are available, the latter two assumptions need not be imposed, and can be tested. The stochastic econometric frontier is estimated several ways in order to test the sensitivity of the results to these assumptions.

The thick frontier model is estimated separately for the highest and lowest cost quartiles. Error terms within quartiles are assumed to represent mean zero, finite variance statistical noise, and to have first-order serial correlation in the panel estimations. Measured inefficiencies are embedded in the difference in predicted costs between the highest and lowest quartiles. This difference may occur in the intercepts or in the slope parameters.

The disturbance terms on the input share equations, ψ_k, follow normal distributions with finite mean and variance for both frontier approaches. However, allowing ψ_k to have a non-zero mean in the stochastic econometric model allows for persistent allocative inefficiency over time (see Bauer, Ferrier, and Lovell, 1987). These parameters are identified because the share equation intercepts remain identified through cross-equation restrictions.

III. Measures of Inefficiency and Total Factor Productivity

Within the stochastic frontier approach, four separate techniques are used to estimate individual bank inefficiencies, although all four use essentially the same cost function shown in equations (1)-(2) above. Two panel estimation techniques, based on Schmidt and Sickles (1984), assume that bank inefficiencies are fixed over time.[6] The GLS panel estimator makes the further assumption that the inefficiency disturbances are uncorrelated with the regressors. The cost equations are estimated, and a separate intercept α_i for each of the 683

6. This assumption is not strictly necessary. Cornwell, Schmidt, and Sickles (1990) generalized the approach to allow inefficiencies to vary over time, but in a structured manner.

banks is recovered as the mean residual for that bank. The most efficient bank in the sample is assumed to be fully efficient, and the inefficiency of bank i is given by the proportionate increase in predicted costs over the efficient bank, or $\hat{\alpha}_i - \min_j \hat{\alpha}_j$. The WITHIN panel estimator uses a fixed-effects model to estimate α_i, where the variables are measured as deviations from individual bank means, eliminating the need to assume that inefficiency is uncorrelated with the regressors. Both of these estimators allow the statistical error terms to be correlated across equations and over time.

Some strong assumptions are required for these techniques to yield consistent estimates of inefficiency. First, inefficiency must be the only time-invariant fixed effect. Second, as the number of banks approaches infinity, the density of the inefficiency disturbances must be non-zero in the neighborhood of $(0,\omega)$ for some $\omega > 0$. That is, as the sample size increases, it must become more likely that firms on the estimated frontier are near the true frontier.

Two MLE estimation techniques are based on Bauer, Ferrier, and Lovell (1987). The composed error ε is the sum of the half-normally distributed inefficiency μ and the normally distributed statistical error ν. The measured inefficiency for an individual bank is the conditional mean $E(\mu|\varepsilon)$.[7] The MLE (by year) technique allows the translog coefficients to vary over time, while the MLE (panel) technique holds the coefficients fixed over time. Unlike the other panel techniques, MLE (panel) does not allow for autocorrelation.

In the thick frontier approach of Berger and Humphrey (1991a,b), the difference in predicted average costs between cost quartiles is decomposed into explained and unexplained parts, and the unexplained part is taken to be the inefficiency difference between the quartiles. The proportionate difference in predicted average costs is given by

$$DIFF = (\hat{AC}^{Q4} - \hat{AC}^{Q1})/\hat{AC}^{Q1}, \qquad (3)$$

where $\hat{AC}^{Qi} = \hat{C}^{Qi}(X^{Qi})/TA^{Qi}$ is predicted average cost, $\hat{C}^{Qi}$ incorporates the parameter

7. The formula is $E(\mu|\varepsilon) = (\sigma_\mu \sigma_\nu/\sigma)\bullet[[\phi(\varepsilon\lambda/\sigma)/[1 - \Phi(\varepsilon\phi/\sigma)]] - [\varepsilon\lambda/\sigma]]$, where σ_μ and σ_ν are the variances of μ and ν respectively, $\lambda \equiv \sigma_\mu/\sigma_\nu$, and $\sigma^2 \equiv \sigma_\mu^2 + \sigma_\nu^2$. Clearly, this approach requires μ and ν to be independent of each other and of the regressors. An alternative is to use the conditional mode, which was found to be nearly the same here.

estimates obtained using the Qi data, and X^{Qi} and TA^{Qi} are the vector of mean regressors and the mean total assets respectively for the size class for the ith quartile (size class scripts are suppressed for expositional ease). It is assumed that differences in output levels, output mix, and input prices are due not to inefficiencies, but to exogenous differences in the markets in which banks operate. The part of DIFF that cannot be attributed to the exogenous variables in the model constitutes the measured inefficiency residual, given by:

$$INEFF = (\hat{AC}^{Q4} - \hat{AC}^{Q4*})/\hat{AC}^{Q4}, \tag{4}$$

where $\hat{AC}^{Q4*} = \hat{C}^{Q1}(X^{Q4})/TA^{Q4}$ is the predicted unit cost for Q4 data evaluated using the "efficient" Q1 technology. Thus, INEFF captures only the unexplained difference in the estimated cost functions, holding the data constant at Q4. Included in INEFF are overpayments of deposit and purchased funds interest, as well as operating cost inefficiencies.

For the thick frontier approach, as for the stochastic econometric approach above, both panel and cross section methods by year are employed. First, a panel estimation is used in which the estimated cost function parameters are held constant over the entire time period, and the average cost quartiles are formed on the basis of costs over the entire period as well (stable cost function, stable quartiles). This is analogous to the stochastic econometric frontier panel models [GLS, WITHIN, and MLE (panel)]. Second, separate cross section estimates are made for each year of the sample, but the quartiles are based on the entire time period (varying cost function, stable quartiles). As discussed below, this is our preferred method because it allows for changes over time in the technology and environment of banks (reflected in changing slope parameters), and yet eliminates much of the year-to-year noise for individual banks when choosing which are most and least efficient. This technique of basing the frontier on the entire time series panel, but allowing the parameters to vary over time, is practical only for the thick frontier approach and has no analog in the stochastic frontier approach. Finally, separate cross section estimates are made for each year in a model in which the average cost quartiles are also formed by year (varying cost function, varying quartiles). This is analogous to the MLE (by year) stochastic econometric model.

We turn next to the measurement of total factor productivity (TFP), which reflects technical change plus the scale economy effects on costs associated with variations in output levels over time. Measurement of the technical change component of TFP in a service industry like banking is difficult. Unfortunately, there is no unique indicator or proxy for measuring the effects of technical change, either for neutral or embodied technical progress. As a consequence, virtually all previous banking studies have chosen to model technical change as a simple time trend. However, studies of electric utilities have suggested that time trends may poorly reflect year-to-year variations in technical change when this process is not constant or smoothly increasing or decreasing (Kopp and Smith, 1983; Nelson, 1986). As a result, we adopt an index approach which allows technical change to vary freely over time. As developed by Caves, Christensen, and Swanson (1981) and Baltagi and Griffin (1988), the index approach is a generalization of Solow's index of technical change A(t).

In the pooled models, time-specific intercept shift variables are specified to reflect neutral technical change. In cost equation (1), the intercept term α is replaced by $\sum_{t=1}^{12} \eta_t D_t$, where D_t equals 1 in period t and 0 otherwise (t = 1,..,12 over 1977-88).[8] The growth rate of technical progress from t to t+1 is the common rate of input reduction, holding outputs fixed:[9]

$$\text{INDEX}_{t+1,t} = -(\partial \ln TC/\partial D_{t+1} - \partial \ln TC/\partial D_t) = -(\eta_{t+1} - \eta_t), \qquad (5)$$

where the negative sign turns cost reductions into technical advances.

8. This is a simplification of Caves, Christensen, and Swanson (1981), who allowed D_t to interact with the regressors. Baltagi and Griffin (1988) extended the Caves et al. specification by imposing a set of nonlinear restrictions on the D_t parameters to obtain the same A(t) effect for neutral, nonneutral, and scale-augmenting technical change. When such nonlinear restrictions were used in Humphrey (forthcoming), estimation was time consuming, and importantly, the technical change estimates were almost identical without these restrictions. Similarly, the interactions of D_t with outputs and input prices did not greatly alter the technical change conclusions here.

9. Technical change can alternatively be expressed as the common rate of output expansion holding inputs fixed. As shown in Caves et al., the two definitions are identical if there are no scale economies. They are close to each other here, since measured scale economies and diseconomies are small.

A more general technique, possible when there is sufficient cross section variation, allows *all* cost function parameters to be affected by technical change, not just intercept shifts (e.g., Berger and Humphrey, 1991b). This technique is equivalent to estimating equations (1)-(2) separately for each annual cross-section time period, and essentially nests the time-specific index technique within it.[10] For the thick frontier approach, the growth rate of technical change is the proportional decline in predicted average costs using the estimated parameters from periods t+1 and t, but evaluated using data only from the base period t:

$$SHIFT_{t+1,t} = -(\hat{AC}^{Q1*}_{t+1} - \hat{AC}^{Q1}_{t})/\hat{AC}^{Q1}_{t}, \qquad (6)$$

where $\hat{AC}^{Q1}_{t}$ is the predicted average cost for thick frontier banks in period t defined above, and $\hat{AC}^{Q1*}_{t+1} = \hat{C}^{t}_{t+1}(X^{Q1}_{t})/TA^{Q1}_{t}$ is the predicted average total cost for the same banks using period t+1 technology. For the stochastic econometric frontier, the average costs in (6) refer to all banks rather than just those on the frontier, since the frontier consists of only one bank per year that might be significantly different from the adjacent years.

Scale economy effects are added to the technical change effects to yield TFP. These effects combine the overall cost elasticity, $SCE = \sum_{i=1}^{5} \partial \ln TC/\partial \ln Y_i$, with the proportional change in the cost share (c_i) weighted average of the five outputs, $\dot{Y} = d\ln(\sum_{i=1}^{5} c_i Y_i)$, all in real terms.[11] As for SHIFT, $\dot{Y}$ uses the lowest cost quartile under the thick frontier approach and the overall average data under the stochastic econometric approach. Thus, TFP is expressed as:[12]

$$TFP = (INDEX \text{ or } SHIFT) + (1 - SCE)\dot{Y}. \qquad (7)$$

Inefficiency can be incorporated into a TFP measure for all banks (Bauer, 1990), but in this

10. One unnested difference between the cross section and pooled techniques is that the unit banking dummy U was deleted from the cross section estimations because of collinearity problems. Another difference is that the cross section estimations do not account for autocorrelation.

11. The c_i cost shares were taken from the Federal Reserve's *Functional Cost Analysis* (FCA) report and reflect the operating and interest expenses allocated to the 5 output categories for the set of large banks in the FCA report.

12. Bauer (1990) incorporated inefficiency in a TFP measure for all firms, but it is excluded here because our TFP measure applies only to frontier banks.

application they are kept separate, as our TFP applies only to frontier banks.[13]

IV. Bank Inefficiency Estimates for 1977-88

Stochastic Econometric Frontier Average Inefficiencies. Columns (1)-(4) of Table 1 show the average inefficiency estimates for the four stochastic econometric frontier techniques, which range from about 7 to 17 percent for the entire time period. Our preferred model is MLE (by year) in column (3), since the measured inefficiency is free to vary by year (unlike GLS and WITHIN), and the cost function parameters are also free to vary [unlike GLS, WITHIN, and MLE (panel)]. In this preferred model, bank inefficiency averages 15 percent.

Despite the flexibility of this approach, the measured variation in inefficiency over time is rather small. The largest variation occurs in the early 1980s, when inefficiency is seen to fall with the advent of deposit interest rate deregulation, i.e., the establishment of new types of consumer accounts and the removal of interest rate ceilings on existing accounts. The measured fall in inefficiency may reflect a temporary disequilibrium in which the most efficient banks were also the most aggressive in raising rates and going after new funds. Examination of the pattern of inefficiencies by size class, shown in Table 2 for the MLE (by year) technique, suggests that larger banks, at 19 percent average inefficiency, may be slightly more inefficient than smaller ones.

Comparison among the stochastic econometric techniques finds about the same levels of average inefficiency for the three techniques other than GLS, raising suspicions about the GLS assumption that inefficiency is uncorrelated with the regressors. However, the correlations among the measures across banks are surprisingly high (not shown in tables). The R^2 between the GLS and WITHIN estimates is .89, and the R^2 between the two MLE methods is .93. The R^2 between each of GLS and WITHIN and each of the MLE methods is lower, between .38 and .50, in part because GLS and WITHIN force inefficiency to be time-invariant.

---

13. Denny, Fuss, and Waverman (1981) attempt to measure other aspects of TFP, such as the effect of deviations of input prices from marginal outlays. However, such aspects are likely to be of little consequence here, since banks are reasonably competitive in most of their input markets.

The assumption of a half-normal distribution for the inefficiencies is examined in Figure 2, which shows a histogram of the inefficiency estimates $E(\mu|\varepsilon)$ for the preferred model, MLE (by year). The shape of this empirical distribution, as well as the distributions for the other three models shown in Figures 3, 4, and 5, appears to be roughly consistent with the half-normal assumption. Previous studies have often used the half-normal assumption, but have not examined its consistency with the data.

Thick Frontier Interquartile Inefficiencies. Columns (5)-(7) of Table 1 show the interquartile inefficiency estimates for the three thick frontier techniques, which range from about 16 to 21 percent for the overall time period. Our preferred model is the varying cost function, stable quartiles model shown in column 6, since all of the frontier parameters are allowed to vary across years, maximizing flexibility, but the cost quartiles are stable, minimizing the effects of temporary or random fluctuations in costs. The results for the preferred model indicate an average interquartile inefficiency of 21 percentage points of the 23 percent average difference in predicted and actual costs. Differences between high cost and low cost banks in their output levels, input prices, and other exogenous variables explain the remaining 2 percentage points. In this model, inefficiency has some year-to-year variation, but no strong upward or downward trend is evident. Again, the largest variation occurs in the early 1980s, when inefficiency falls with the advent of deposit interest rate deregulation.

A breakout of the inefficiencies by size class, shown in Table 2, suggests no particular trend except that banks in the largest size class (assets > $10 billion) have more than 48 percent inefficiencies, substantially greater than the other size classes and exceeding the actual or predicted cost differences for this size class. This suggests that the cost function parameters, which are dominated by the observations on smaller banks, may not extrapolate well to the relatively few large banks. If the largest size class is deleted, average interquartile inefficiency is reduced from 21 to 17 percent.

As expected, the model in column 7 in which both the frontier parameters and the cost quartiles are allowed to vary over time shows the greatest year-to-year variation in the inefficiency estimates. However, the average results are very similar to those for the preferred model, 20.8 versus 21.0 percent average inefficiency. In contrast, when both the frontier

parameters and the quartiles are held constant over the entire period in column 5, the inefficiency estimates have a downward trend starting in the early 1980s. Average inefficiency for this model is 15.7 percent, significantly lower than for the other thick frontier models.

Comparison of Stochastic Econometric and Thick Frontier Inefficiencies. As noted above, the measured inefficiencies of the stochastic econometric and thick frontier approaches are not strictly comparable because the former takes the average comparison of all banks to the frontier, while the latter compares the average bank in two different quartiles. Table 3, however, transforms these approaches into comparable forms and contrasts the preferred MLE (by year) stochastic econometric approach with the preferred Varying Function, Stable Quartiles thick frontier approach. Interquartile inefficiencies are first computed for both approaches using quartiles based on the inefficiencies estimated using the stochastic econometric frontier approach (columns 1 and 2). The procedure is then repeated using actual average costs to form the quartiles, i.e., using the thick frontier quartiles. In all cases, the data are stratified by size class before the quartiles are obtained.

When each method uses quartiles based on its own approach, the stochastic econometric and thick frontier approaches yield similar interquartile inefficiencies of 18.4 and 21.0 percent in columns (1) and (4) respectively. These estimated inefficiencies become 4.7 and 27.7 percent in columns (3) and (2) respectively when the other method is used to compute the quartiles, suggesting that there are important differences in the bank inefficiency rankings generated by the two approaches. Further examination reveals that of the banks identified as being in the most efficient quartile in one method, only 38 percent are also identified as being in the most efficient quartile in the other method. If the two methods were totally unrelated, 25 percent would be expected to match. The matching percentage for the least efficient quartiles is 46 percent.[14] Thus, the data suggest that while the two methods find nearly the same level of inefficiencies, there are important differences between them in terms of which banks are identified as being the most and least efficient.

14. The higher matching percentage for the least efficient banks may reflect the skewed nature of the inefficiencies, shown in Figure 2. The least efficient banks have much higher costs than other banks and may be easier to identify.

Comparison with Other Studies. The findings here of inefficiencies on the order of 15 to 21 percent of costs are similar to those found in the extant bank efficiency literature, although as noted, the results are not always strictly comparable. One of the stochastic econometric studies of banks, Ferrier and Lovell (1990), essentially applied the MLE (by year) method to a single year of data. They found average inefficiencies of 26 percent for a sample of small to medium sized banks for 1984. The similarity to the findings here is somewhat surprising, given the many differences between the studies. Ferrier and Lovell used smaller banks, used a different definition of bank output (number of accounts instead of dollar values), and excluded interest expenses, which make up the majority of bank costs. Our results are at the low end of Berger's (1991) range of about 10 percent to several hundred percent inefficiency for samples of all sizes of banks in the 1980s. Berger applied techniques similar to the GLS and WITHIN frontier methods, but with parameters that vary by year and with some truncation of outliers. Our results are also within the range of findings of the previous thick frontier models of banking by Berger and Humphrey (1991a,b), who found average interquartile inefficiencies of between 17 and 42 percent when examining banks of all sizes in the 1980s.[15]

V. Total Factor Productivity and Scale Economies for Banks for 1977-88

Total factor productivity combines the effects of technical change and changes in scale as output expands over time. Estimates of TFP for the four stochastic econometric frontier methods and the three thick frontier methods are shown in the top panel of Table 4. Negative growth rates are obtained for six of the seven estimations. These range from -3.55 percent to +0.16 percent annually and represent a striking effect of unusual changes in the banking industry over this time period. Because the scale economy estimates are so close to constant

15. DEA frontier studies of banking find average inefficiencies of (in ascending order of magnitude) 12 percent by Elyasiani and Mehdian (1990), 21 percent by Ferrier and Lovell (1990), 43 percent by Rangan, et al. (1988), 54 percent by Aly, et al. (1990), and 70 to 105 percent by Ferrier et al. (1990). Our results may be lower than most of these because of the upward bias in DEA from counting all random error as inefficiency.

average costs (shown below), the TFP estimates almost exclusively reflect technical change.[16]

Although the negative TFP estimates are surprising, they are consistent with a number of other studies of bank TFP and technical change during this period. Negative technical change is found (a) when all banks in our panel are used (instead of only frontier banks) and (b) when technical change is alternatively represented by a time trend, a cost curve shift, or a more comprehensive set of time-specific shift variables than is specified here (see Humphrey, forthcoming). Negative to small positive TFP growth rates were also found using aggregate bank data in a growth accounting model and in an estimated cost function over 1967-87 (Humphrey, 1991). While some studies of bank technical change, Hunter and Timme (1986), Evanoff, Israilevich, and Merris (1989), and Hunter and Timme (1991), have reported larger positive growth rates, the underlying explanation may be methodological differences.[17]

There are several possible explanations for the measured poor productivity growth of banks over this time period. In the late 1970s, historically high interest rates greatly increased the use of cash management techniques by corporations. This reduced demand deposit balances, which did not pay explicit interest, and forced banks to rely more heavily on higher-cost funds.[18] Such an increase in real costs is measured as a reduction in TFP.

The increased interest costs from corporate cash management were extended to consumer deposit accounts with the deregulation of the early 1980s. Depositors were able to shift noninterest-earning demand deposits into interest-earning checking plans (NOW accounts) beginning in 1981, and were able to shift into variable-rate Money Market Deposit

---

16. To illustrate that TFP and technical change are nearly identical, the annual average technical change underlying the -0.39 percent, -2.28 percent, and -2.14 percent thick frontier TFP annual growth rates in Table 4 are -0.30 percent, -2.13 percent, and -1.97 percent respectively. The almost negligible effect of scale economies on TFP suggests that the use of alternative indices of the change in output (e.g., the Tornquist-Theil discrete approximation to a continuous Divisia index) would have little effect on the results.

17. The first two studies cited used only operating costs, which reflect only around 25 percent of the total costs used here, and thus may not indicate technical change for the entire banking operation. The latter study used total costs, but contained a specification difficulty that, once adjusted for, turned their positive growth rate to negative (see Humphrey, forthcoming, for details).

18. See Porter, Simpson, and Mauskopf (1979) for a description of this process.

Accounts (MMDAs) by 1983. Interest rate ceilings on all other deposits were phased out by 1986 as well. Competition among banks increased as regulatory impediments to such competition were reduced, raising bank costs and contributing to negative measured TFP growth.

Increased competition from outside of banking also increased during this time period, raising banks' costs of funds. Thrift institutions were given greater powers to compete for consumer funds, particularly the ability to offer checkable deposits, reducing the market power of banks. Similarly, nontraditional sources of competition, such as money market mutual funds that sold shares in portfolios of short-term Treasury securities, provided alternatives to federally insured deposits.

Thus, over the late 1970s and the early 1980s, banks lost much of their monopsony power over their depositors, in part due to the actions of their corporate customers, in part due to the deregulation of consumer deposit rates, and in part due to increased nonbank competition. In all cases, banks' costs were driven up and measured TFP was driven down. Berger and Humphrey (1991b) estimated that as a net result of these changes, aggregate bank profits earned through the payment of below-market rates on deposits fell from $61 billion in 1980 to $4 billion in 1988 (in constant 1988 dollars).

It might have been possible for banks to offset these negative TFP factors by lowering operating costs, especially by closing branches. Indeed, a major technical innovation of the period, automated teller machines (ATMs), was predicted to facilitate the closing of many branches. However, to the contrary, the number of bank branches actually increased in the 1980s.[19] Part of the reason appears to be that the increased competition for depositors forced banks to provide convenient branches and ATMs for consumers, as well as higher interest rates. According to industry surveys, choice of bank by depositors is largely based on convenience. Part of the reason may also relate to enforcement of the Community Reinvestment Act, which encouraged banks to keep open some uneconomic branches in certain local communities that might otherwise have been closed. In addition, the benefits of ATMs may have

19. Over the decade, banks closed about 6,650 branches, but opened approximately 16,500.

largely been captured by consumers, just as were the benefits of deposit rate deregulation. While the average cost of a single ATM transaction may be substantially less than that of using a human teller, the added convenience of ATMs appears to increase the number of transactions substantially. For example, customers may withdraw less cash during a typical ATM transaction than during a typical human teller transaction, which increases the total number of transactions and operating costs absorbed by the bank (see Berger, 1985).

This analysis may explain why researchers have failed to observe much positive technical change or productivity growth in banking during the last one and one-half decades. All of the important changes described here, cash management improvements, deregulation of deposit rates, increased nonbank competition, and the ATM innovation, in principle should have increased productivity in the banking sector, but not necessarily in its measured component. While measured productivity growth has been nonexistent, the users of banking services have benefited from higher deposit interest rates, added convenience of ATMs, and an increased number of branches. These benefits, which constitute increases in the "quality" of banking services, are not captured in any measure of banking output. Thus, although there has been no measured productivity growth, it would be inappropriate to conclude that society as a whole has not benefited. Rather, there has been a substantial redistribution of productivity benefits in which users of banking services have gained at the expense of banks.[20]

We turn finally to examination of the scale economy component of TFP, which has often been considered to be an important topic in banking of its own merit. Since most studies of bank scale economies have focused on smaller banks, it may be of particular interest to investigate the scale economies of the 12 annual samples of relatively large banks studied here. The scale economies derived from the seven stochastic econometric and thick frontier models are shown in Table 4. The figures for the individual years are multiproduct

20. An analogous situation occurred in the electric utility industry during the 1970s, when expensive pollution control restrictions were mandated. The measured output of this industry, kilowatt hours, did not rise commensurately with the increased costs, so that measured TFP fell (see Gollop and Roberts, 1983). However, society may still have benefited on net through improvements in air quality, but these are not incorporated in measured industry output.

ray scale economies, $\sum_{i=1}^{5} \partial \ln TC/\partial \ln Y_i$, averaged across size classes. A figure less than or greater than 1 indicates scale economies or diseconomies respectively.

The estimates vary across estimation method from slight economies of about 5 percent to slight diseconomies of about 4 percent (i.e., $.95 \leq \sum_{1} \partial \ln TC/\partial \ln Y_i \leq 1.04$). These results are also quite stable over time. Unlike the inefficiency and TFP results, the deregulation of the early 1980s does not appear to have affected scale economies significantly. A breakout by size class, shown in Table 5 for the preferred stochastic econometric and thick econometric models, indicates some minor variation by size of bank. For the preferred stochastic econometric model, every size class shows scale diseconomies of 1 to 3 percent on average, except that the largest size class (assets > $10 billion) has average diseconomies of 5 percent. For the preferred thick frontier model, the scale diseconomies fall from an average of 8 percent for the smallest size class ($100 million ≤ assets < $200 million) to approximately constant average costs for the top two size classes (assets > $5 billion).

The small scale economy and diseconomy estimates found here both on and off the frontier are consistent with most of the conventional studies of bank scale economies (see the surveys by Mester [1987], Clark [1988], and Humphrey [1990]).[21] An earlier study that compared frontier and nonfrontier scale economies (Berger and Humphrey, 1991a) also found little difference, suggesting that the economies found here may well represent the universe of banks in branching states with assets over $100 million, rather than just the relatively efficient ones. An additional conclusion is that the average scale economies and diseconomies of about 5 percent or less found here and elsewhere appear to be dominated by inefficiencies, which average about 15 to 20 percent here and are higher in some other studies.[22]

---

21. Studies finding larger scale economies typically have measured how operating expenses, rather than total costs, vary with bank scale. The use of operating costs alone tends to bias the results toward finding scale economies, since banks generally substitute interest-cost-intensive purchased funds for operating-cost-intensive produced deposits as they increase scale, making operating costs per unit of output decline without any real basis.

22. The ray scale economy measure used here is a local, rather than global, concept, and thus it could understate the gains to scale when exceptionally large changes in scale are involved (see Evanoff and Israilevich, 1991). However, ray scale economies fairly accurately portray the cost effects of the changes in size that actually occur. Moreover,

(Footnote continues on next page)

VI. Conclusions

This paper compares two general approaches to estimating inefficiency in banking, the stochastic econometric approach and the thick frontier approach, as well as examining several specific techniques within each approach. We also employ these methods to obtain estimates of productivity growth and scale economies in the banking industry. The data set to which the analysis is applied is a panel of 683 large U.S. branching state banks that account for two-thirds of all U.S. banking assets. The data cover 1977-88, a period of significant financial market innovation, deregulation, and technical innovation in banking.

The levels of bank inefficiency found here are reasonably consistent between the two approaches. Using the preferred models of each of the stochastic econometric and thick frontier approaches, the average difference in efficiency between the most and least efficient quartiles of banks is estimated to be 18 and 21 percent respectively. Similarly, the average efficiency of all banks is estimated to be 15 percent using the preferred technique of the stochastic econometric approach. However, while the two approaches yield similar average efficiency findings, they rank individual banks quite differently.

The inefficiency estimates found here are consistent with those in the extant bank inefficiency literature, but are toward the low end of the literature's estimates. Nevertheless, these inefficiencies are sufficiently large to dominate the scale economy effects of 5 percent or less found here and elsewhere in the literature. This finding suggests that analyses which focus on the scale of bank operations may be misplaced. Further increases in competition in the banking industry are more likely to put pressure on inefficient banks of all sizes than to force banks of any particular size to exit the industry.

(Footnote continued from previous page)
the relatively flat average cost curves shown in Figure 1A, where the AC_{MEAN} curve varies by only 5 percent across all size classes, suggest that even very large changes in scale are not associated with large changes in average costs. In addition, Berger (1991), the only study to compute both conventional efficiencies and scale efficiencies (comparisons of average costs for each bank to those for the scale-efficient bank of the same product mix), found inefficiencies to dominate scale effects.

The stochastic econometric frontier and thick frontier approaches also give similar estimates for TFP growth and its two components--technical change and scale economies. Estimates of annual TFP growth ranged from negative to small positive values, from -3.55 percent growth per year to +0.16 percent. These surprising results, which at first blush suggest technical retrogression, appear to be consistent with some institutional events that occurred over this time period. In particular, over the late 1970s and the early 1980s, deposit interest costs rose sharply as banks lost much of their monopsony power over their depositors, which had allowed them to pay below-market rates. This loss, which was the depositors' gain, was due to more sophisticated corporate cash management techniques, the deregulation of consumer deposit rates, and an increase in nonbank competition. The higher cost of funds is measured as a negative technical change because costs increased without a corresponding increase in measured output. The benefits of the key technical innovation of the period, ATMs, also appear to have been largely captured by consumers, who enjoyed more convenient service without paying significantly more for it. Thus, despite the fact that *measured* productivity fell, the unmeasured extra product of the industry in the form of more favorable deposit rates and more convenient transactions for depositors implies that the true productivity of this industry may well have increased.

Turning to future implications, forthcoming increases in competitive pressure in banking will most likely come from bank mergers, both within and across markets. Several large banking organizations are in the process of, or have already completed, in-market horizontal mergers. In addition, if interstate banking legislation passes, substantially greater opportunities for across-market mergers will be created. In these next rounds of increased competition, there will be considerably less room for depositor benefits than in the previous rounds, since most banks pay close to market rates already. However, the substantial inefficiencies cited above leave room for some *measured* increases in bank productivity if efficient banks take over inefficient banks and raise the efficiency of the latter group significantly.

REFERENCES

Aigner, D., C.A.K. Lovell, and P. Schmidt, "Formulation and Estimation of Stochastic Frontier Production Function Models," *Journal of Econometrics*, 86 (1977): 21-37.

Aly, H.Y., R. Grabowski, C. Pasurka, and N. Rangan, "Technical, Scale, and Allocative Efficiencies in U.S. Banking: An Empirical Investigation," *Review of Economics and Statistics*, 72 (1990): 211-18.

Baltagi, B.H., and J.M. Griffin, "A General Index of Technical Change," *Journal of Political Economy*, 96 (1988): 20-41.

Battese, G.E., and G.S. Corra, "Estimation of a Production Frontier Model with Application to the Pastoral Zone of Eastern Australia," *Australian Journal of Agricultural Economics*, 21 (1977): 169-179.

Bauer, P.W., "Decomposing TFP Growth in the Presence of Cost Inefficiency, Nonconstant Returns to Scale, and Technological Progress," *Journal of Productivity Analysis*, 1 (1990), 287-99.

Bauer, P.W., G. Ferrier, and C.A.K. Lovell, "A Technique for Estimating a Cost System that Allows for Inefficiency," Working Paper, Federal Reserve Bank of Cleveland (1987).

Berger, A.N., "The Economics of Electronic Funds Transfers," Outline, Board of Governors of the Federal Reserve System, Washington, DC (1985).

Berger, A.N., "X-Efficiency and Scale Efficiency in the U.S. Banking Industry," working paper, Board of Governors of the Federal Reserve System, Washington, DC (1991).

Berger, A.N., and D.B. Humphrey, "The Dominance of Inefficiencies over Scale and Product Mix Economies in Banking," *Journal of Monetary Economics* 28 (1991a): 117-48.

Berger, A.N., and D.B. Humphrey, "Measurement and Efficiency Issues in Commercial Banking," in Zvi Griliches, ed., *Output Measurement in the Services Sector*, National Bureau of Economic Research, University of Chicago Press (Chicago, IL), 1991b.

Berger, A.N., G.A. Hanweck, and D.B. Humphrey, "Competitive Viability in Banking: Scale, Scope, and Product Mix Economies," *Journal of Monetary Economics* 20 (1987): 501-520.

Board of Governors of the Federal Reserve System, *Reports of Condition and Income* and *Functional Cost Analysis*, Washington, DC, various years.

Bureau of Labor Statistics, U.S. Department of Labor, *Productivity Measures for Selected Industries and Government Services*, Washington, DC, Bulletin 2322 (1989): 170.

Caves, D.W., L.R. Christensen, and J.A. Swanson, "Productivity Growth, Scale Economies, and Capacity Utilization in U.S. Railroads, 1955-74," *American Economic Review*, 71 (1981): 994-1002.

Clark, J., "Economies of Scale and Scope at Depository Financial Institutions: A Review of the Literature," Federal Reserve Bank of Kansas City *Economic Review*, 73 (1988), 16-33.

Cornwell, C., P. Schmidt, and R.C. Sickles, "Production Frontiers with Cross-Sectional and Time-Series Variation in Efficiency Levels," *Journal of Econometrics*, 46 (1990), 185-200.

Denny, M., M. Fuss, and L. Waverman, "The Measurement and Interpretation of Total Factor Productivity in Regulated Industries with an Application to Canadian Telecommunications," in T.C. Cowing and R.E. Stevenson, eds., *Productivity Measurement in Regulated Industries*, Academic Press, New York (1981): 191-202.

Elyasiani, E., and S.M. Mehdian, "A Non-Parametric Approach to Measurement of Efficiency and Technological Change: The Case of Large U.S. Commercial Banks," *Journal of Financial Services Research*, 4 (1990): 157-68.

Evanoff, D.D. and P.R. Israilevich, "Scale Elasticity versus Scale Efficiency," *Issues in Financial Regulation*, Federal Reserve Bank of Chicago (1991).

Evanoff, D.D., P.R. Israilevich, and R.C. Merris, "Technical Change, Regulation, and Economies of Scale for Large Commercial Banks: An Application of a Modified Version of Shephard's Lemma," Working Paper, Federal Reserve Bank of Chicago (1989).

F.W. Dodge Division, *Dodge Construction Potentials Bulletin*, Summary of Construction Contracts for New Addition and Major Alteration Projects, New York: McGraw Hill.

Ferrier, G.D., S. Grosskopf, K. Hayes, and S. Yaisawarng, "Economies of Diversification in the Banking Industry: A Linear Programming Approach," Working Paper, Southern Methodist University, Dallas, TX (1990).

Ferrier, G.D., and C.A.K. Lovell, "Measuring Cost Efficiency in Banking: Econometric and Linear Programming Evidence," *Journal of Econometrics*, 46 (1990): 229-45.

Gollop, F.M., and M.J. Roberts, "Environmental Regulations and Productivity Growth: The Case of Fossil-Fueled Electric Power Generation," *Journal of Political Economy*, 91 (1983): 654-74.

Humphrey, D.B., "Why Do Estimates of Bank Scale Economies Differ?" Federal Reserve Bank of Richmond *Economic Review*, 76 (1990): 38-50.

Humphrey, D.B., "Flow Versus Stock Indicators of Banking Output: Effects on Productivity and Scale Economy Measurement," working paper, Federal Reserve Bank of Richmond (1991).

Humphrey, D.B., "Cost and Technical Change: Effects of Bank Deregulation," *Journal of Productivity Analysis* (forthcoming).

Hunter, W.C., and S.G. Timme, "Technical Change, Organizational Form, and the Structure of Bank Productivity," *Journal of Money, Credit and Banking*, 18 (1986): 152-66.

Hunter, W.C., and S.G. Timme, "Technological Change in Large U.S. Commercial Banks," *Journal of Business*, 64 (1991): 339-62.

Hunter, W.C., S.G. Timme, and W.K. Yang, "An Examination of Cost Subadditivity and Multiproduct Production in Large U.S. Banks," *Journal of Money, Credit and Banking*, 22 (1990): 504-25.

Kopp, R., and V.K. Smith, "An Evaluation of Alternative Indices of Technological Change," *Scandinavian Journal of Economics*, 85 (1983), 127-46.

Meeusen, W., and J. van den Broeck, "Efficiency Estimation from Cobb-Douglas Production Functions with Composed Error," *International Economic Review*, 18 (1977): 435-44.

Mester, L.J., "Efficient Production of Financial Services: Scale and Scope Economies," Federal Reserve Bank of Philadelphia *Economic Review*, 73 (1987): 15-25.

Nelson, R.A., "Capital Vintage, Time Trends, and Technical Change in the Electric Power Industry," *Southern Economic Journal*, 53 (1986): 315-32.

Porter, R., T. Simpson, and E. Mauskopf, "Financial Innovation and the Monetary Aggregates," *Brookings Papers on Economic Activity*, 1 (1979), The Brookings Institution, Washington, DC, 213-229.

Rangan, N., R. Grabowski, H. Aly, and C. Pasurka, "The Technical Efficiency of U.S. Banks," *Economics Letters*, 28 (1988): 169-175.

Schmidt, P., and R.C. Sickles, "Production Frontiers and Panel Data," *Journal of Business and Economic Statistics*, 2 (1984): 367-374.

Sickles, R.C., D. Good, and R.L. Johnson, "Allocative Distortions and the Regulatory Transition of the U.S. Airline Industry," *Journal of Econometrics*, 33 (1986): 143-163.

FIGURE 1A

Average Costs by
Bank Asset Size Class and Cost Quartile
(averages over 1977-88; M = million; B = billion; 683 panel banks)

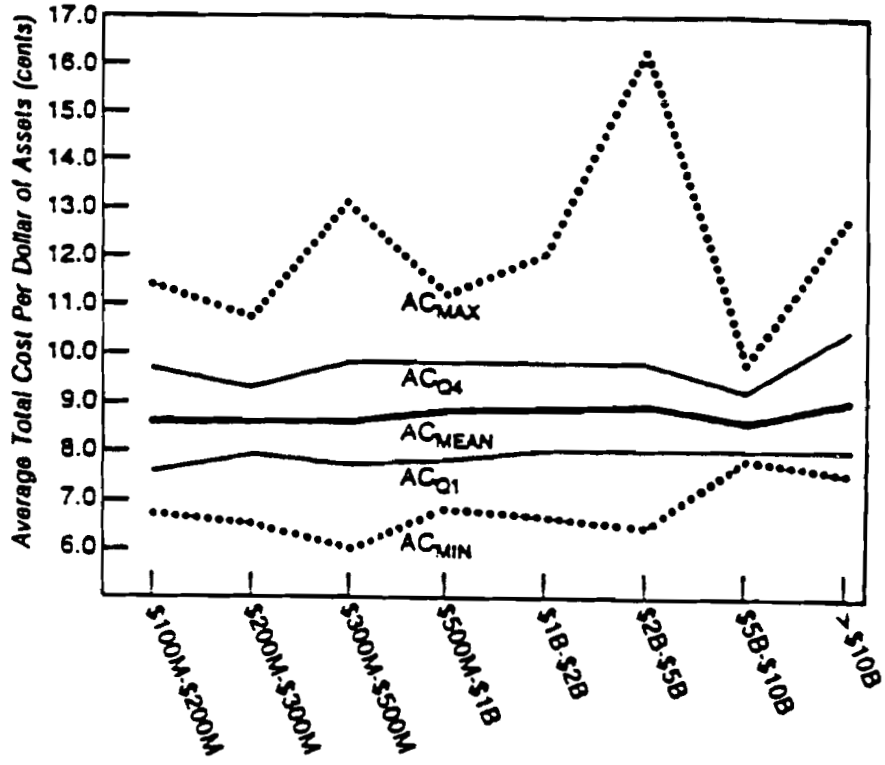
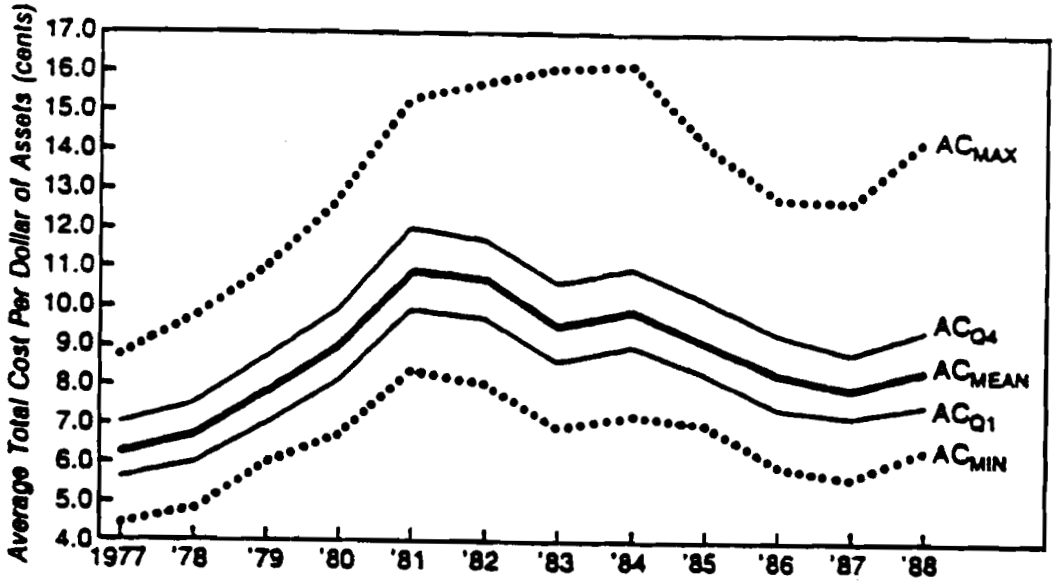


FIGURE 1B

Average Costs over Time and by Cost Quartile
(averages over size classes; M = million; B = billion; 683 panel banks



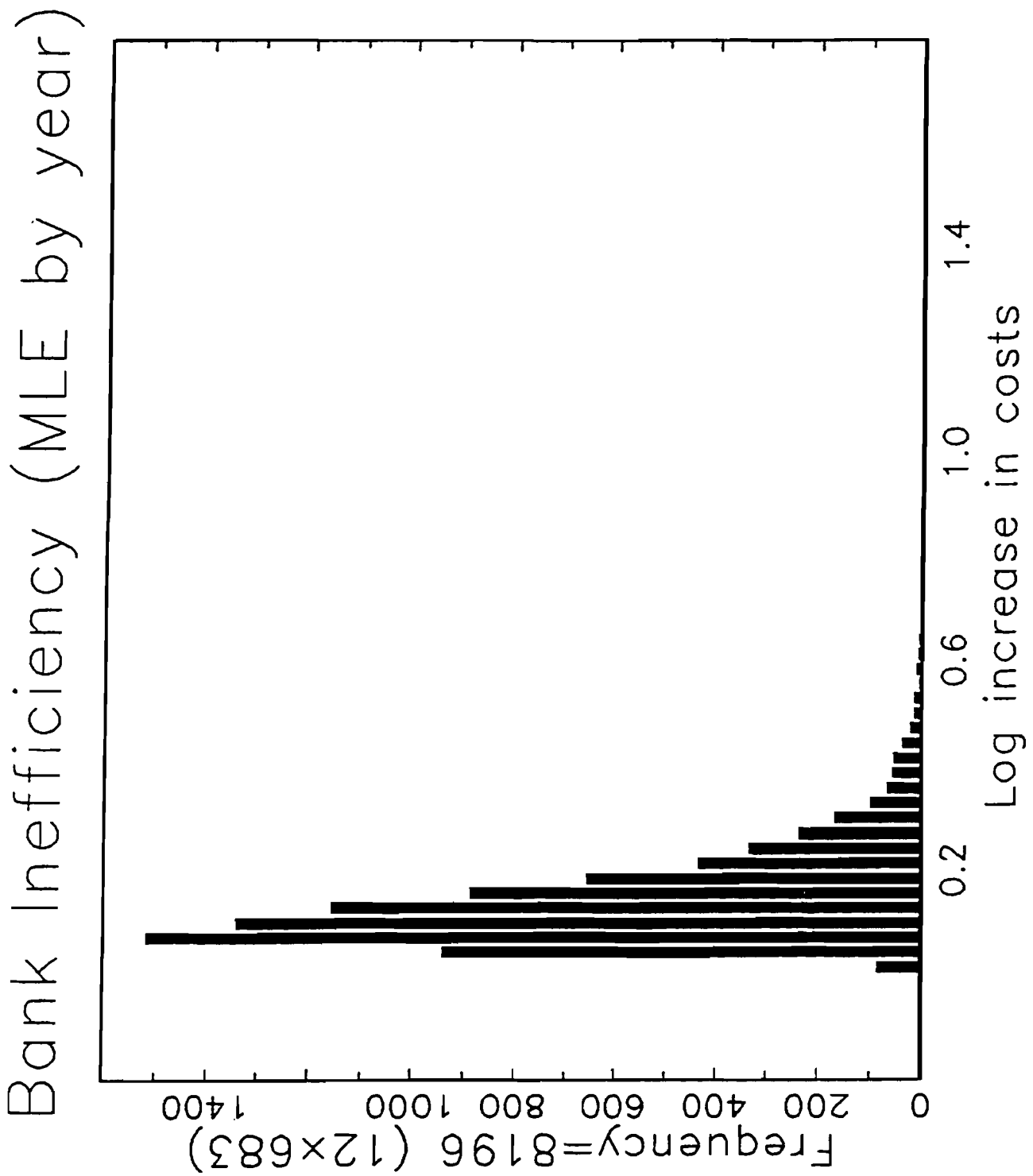


FIGURE 2

FIGURE 3

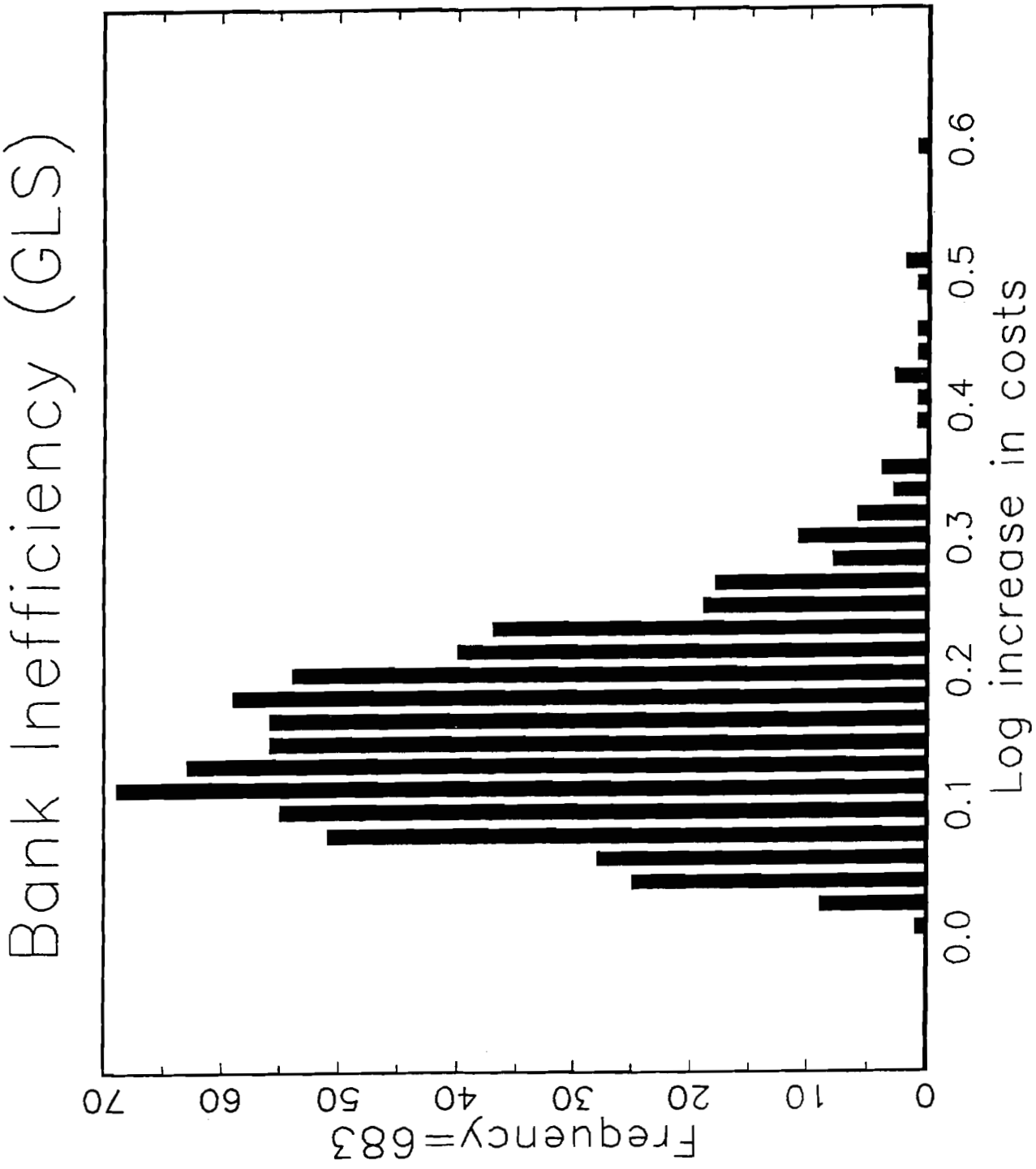
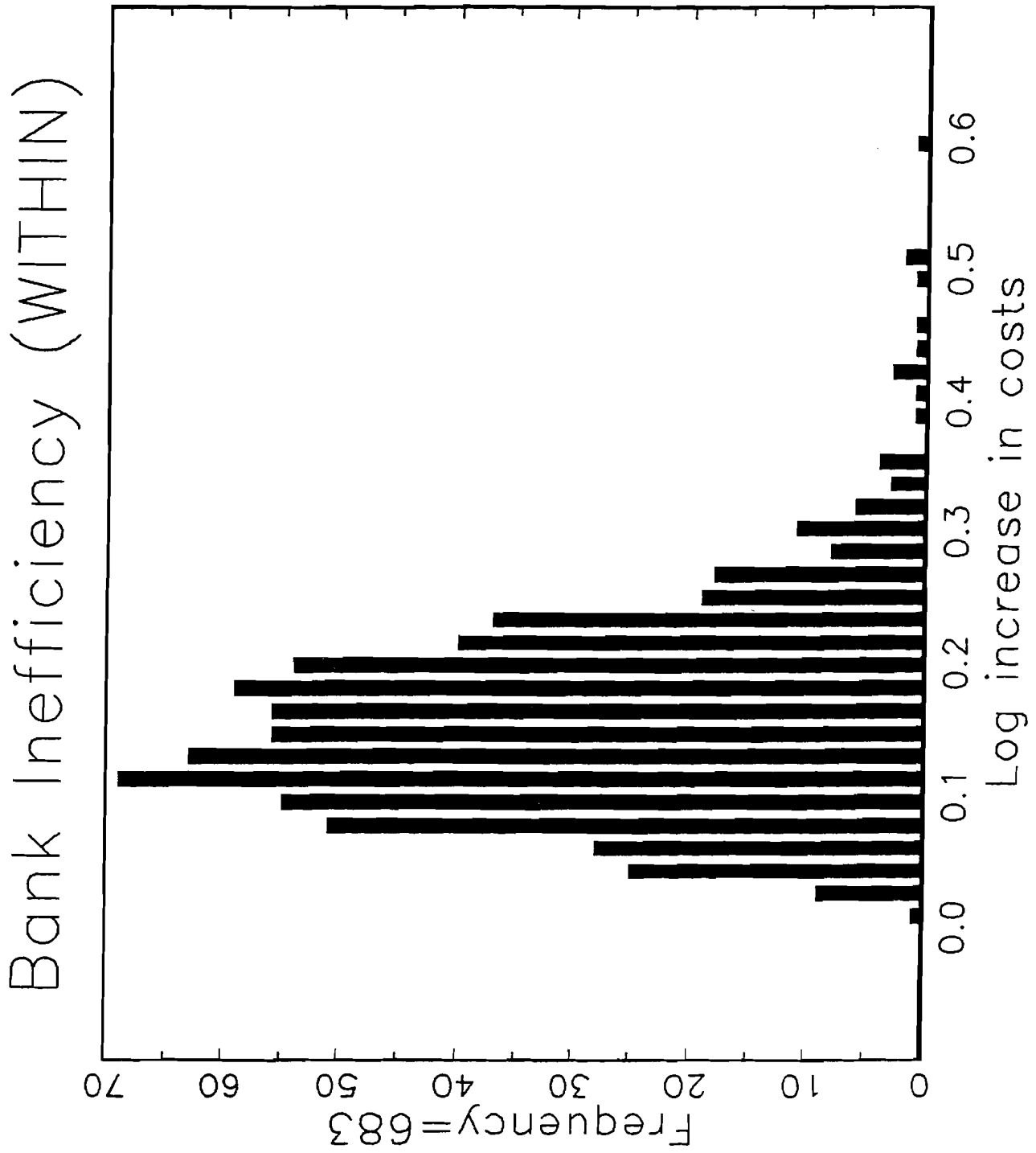


FIGURE 4



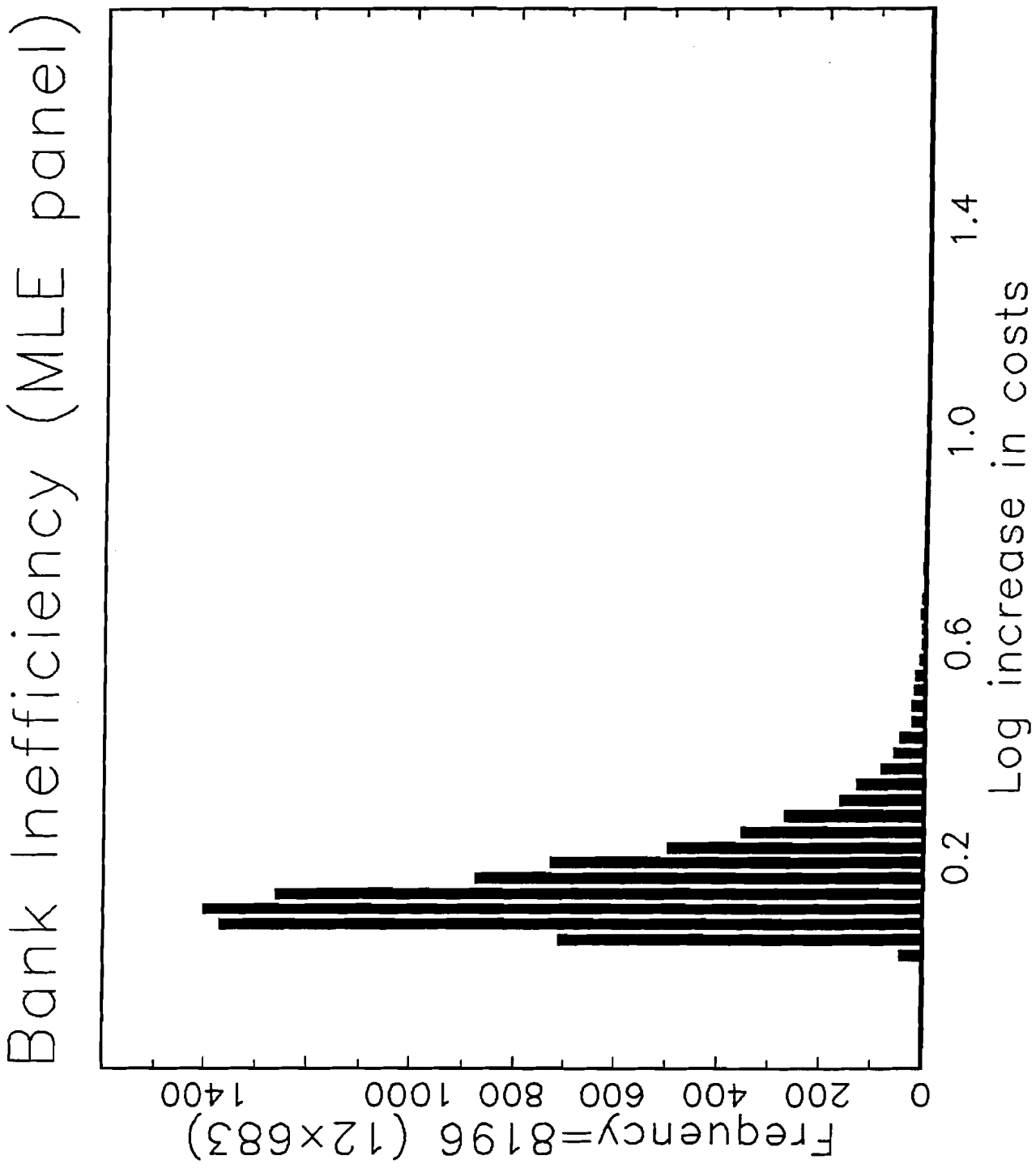


FIGURE 5

Tabl

Estimated Inefficiency for the Stochastic Econometric and Thick Frontier Approaches by Year

	Stochastic Econometric Frontier Approach Average Inefficiencies				Thick Frontier Approach Interquartile Inefficiency Differences (Averaged across size classes)		
	GLS	WITHIN	MLE (by year)	MLE (panel)	Stable Function Stable Quartiles	Varying Function Stable Quartiles	Varying Function Varying Quartiles
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
1977	7.2%	17.4%	15.0%	16.6%	19.0%	16.0%	12.9%
1978	7.2	17.4	16.2	17.0	21.4	20.8	24.4
1979	7.2	17.4	16.1	16.6	23.0	23.7	31.4
1980	7.2	17.4	15.3	16.6	22.4	24.8	30.8
1981	7.2	17.4	16.8	17.2	19.8	19.3	25.8
1982	7.2	17.4	13.5	16.4	15.1	19.3	10.7
1983	7.2	17.4	12.5	15.6	12.5	16.3	5.0
1984	7.2	17.4	15.0	16.0	12.4	19.8	14.7
1985	7.2	17.4	14.3	16.1	11.2	22.7	16.1
1986	7.2	17.4	14.5	16.5	10.7	22.0	15.3
1987	7.2	17.4	17.2	17.5	10.2	21.2	29.3
1988	7.2	17.4	17.2	17.8	10.7	26.5	33.0
Average	7.2%	17.4%	15.3%	16.6%	15.7%	21.0%	20.8%

Notes: For the stochastic econometric approach average inefficiencies, the GLS and WITHIN estimators are based on bank cost function intercepts and assume that bank inefficiency is fixed over time, while the MLE estimators are based on $E(\mu|\varepsilon)$ and allow inefficiency to vary over time.

For the thick frontier approach interquartile inefficiency differences, the methods differ by whether the cost function parameters are assumed to be stable or vary over time, and by whether the average cost quartiles are based on average costs for the entire sample (Stable) or for each year separately (Varying).

Table 2

Inefficiency by Bank Size Class for the Preferred Models of the Stochastic Econometric and Thick Frontier Approaches by Year

Stochastic Econometric Approach Inefficiencies
[MLE (by year)]

Bank Asset Size Classes
(M = million; B = billion)

	100M-200M	200M-300M	300M-500M	500M-1B	1B-2B	2B-5B	5B-10B	>10B	Overall
1977	12.9%	13.8%	15.7%	15.5%	14.3%	15.9%	13.9%	20.5%	15.0%
1978	15.4	15.2	16.2	16.6	16.5	16.3	17.1	20.4	16.2
1979	15.6	14.9	16.2	16.1	16.7	16.7	17.9	17.8	16.1
1980	13.2	14.0	15.4	15.3	15.3	16.0	18.8	16.9	15.3
1981	14.6	15.1	16.9	17.1	16.9	19.1	20.0	18.0	16.8
1982	12.2	12.1	13.3	14.5	14.5	14.8	14.6	14.6	13.5
1983	12.2	11.5	12.0	12.5	13.9	13.2	13.2	16.1	12.5
1984	13.8	13.1	14.8	14.6	16.2	17.1	17.9	22.4	15.0
1985	14.4	12.5	13.9	14.6	15.4	15.8	16.2	18.8	14.3
1986	13.7	12.2	14.4	15.1	15.3	16.3	15.5	17.8	14.5
1987	14.5	13.6	17.9	19.4	19.3	17.9	16.3	21.1	17.2
1988	14.1	14.2	17.5	19.1	19.1	18.4	15.9	23.1	17.2
Overall	13.9%	13.5%	15.3%	15.9%	16.1%	16.4%	16.4%	19.0%	15.3%

Thick Frontier Approach Inefficiencies
[Varying Cost Function, Stable Quartiles]

	100M-200M	200M-300M	300M-500M	500M-1B	1B-2B	2B-5B	5B-10B	>10B	Overall
1977	13.0%	1.0%	11.3%	12.9%	20.0%	16.6%	21.8%	31.1%	16.0%
1978	16.8	8.6	9.9	8.7	13.4	21.8	27.2	59.6	20.8
1979	18.6	10.6	12.7	8.0	13.8	24.6	30.9	70.2	23.7
1980	22.3	10.0	12.0	4.1	12.3	24.2	31.1	82.7	24.8
1981	15.5	12.0	11.7	9.4	13.8	17.6	20.5	54.1	19.3
1982	17.9	14.4	14.2	12.8	14.9	19.4	18.9	41.9	19.3
1983	19.7	15.1	16.2	13.6	15.9	16.8	13.6	19.8	16.3
1984	20.4	15.6	15.9	12.9	16.1	16.4	18.6	42.3	19.8
1985	20.4	18.3	18.9	18.5	16.4	17.7	19.1	52.1	22.7
1986	20.5	18.2	20.4	18.8	17.5	18.6	19.0	42.6	22.0
1987	22.7	18.0	22.3	19.3	17.5	19.4	16.9	33.3	21.2
1988	21.2	21.2	25.0	23.5	20.8	24.4	25.1	51.1	26.5
Overall	19.1%	13.6%	15.9%	13.5%	16.0%	19.8%	21.9%	48.4%	21.0%

Table 3

Interquartile Inefficiency for Both Frontier Methods Using Stochastic Econometric Inefficiencies and Average Costs to Define the Quartiles

	Stochastic Econometric Inefficiencies Define the Quartiles		Average Total Costs Define the Quartiles	
	Stochastic Econometric (1)	Thick Frontier (2)	Stochastic Econometric (3)	Thick Frontier (4)
1977	16.6%	26.7%	8.0%	16.0%
1978	20.0	30.8	7.2	20.8
1979	20.8	34.4	5.3	23.7
1980	18.4	33.9	3.2	24.8
1981	20.1	33.8	2.6	19.3
1982	14.8	28.6	2.5	19.3
1983	14.9	25.0	3.8	16.3
1984	20.6	25.2	4.0	19.8
1985	18.8	24.8	4.6	22.7
1986	17.2	23.0	5.0	22.0
1987	19.6	23.5	5.2	21.2
1988	18.3	22.8	5.3	26.5
Average	18.4%	27.7%	4.7%	21.0%

Notes: The preferred models are used for both approaches, i.e., MLE (by year) for the stochastic econometric approach, and varying cost function, stable quartiles for the thick frontier approach.

Table 4

Total Factor Productivity and Scale Economies for the Stochastic Econometric and Thick Frontier Approaches by Year

	Stochastic Econometric Frontier Approach				Thick Frontier Approach		
	GLS	WITHIN	MLE (by year)	MLE (panel)	Stable Function Stable Quartiles	Varying Function Stable Quartiles	Varying Function Varying Quartiles
				Indices of Total Factor Productivity (1977 = 100)			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
1977	100.0	100.0	100.0	100.0	100.0	100.0	100.0
1978	101.9	90.2	103.5	103.8	97.8	97.8	101.3
1979	102.9	83.8	107.3	105.0	98.3	95.2	102.8
1980	98.5	75.6	103.7	101.7	94.2	91.0	98.7
1981	94.3	69.2	98.1	97.0	90.7	84.0	88.2
1982	91.0	65.4	96.5	89.9	90.1	77.3	75.5
1983	90.3	64.5	100.2	88.4	96.2	78.9	75.2
1984	91.2	64.6	108.2	91.2	92.7	78.3	75.4
1985	91.3	64.9	104.2	89.1	91.7	78.7	74.9
1986	92.5	66.0	102.5	89.2	92.8	80.3	78.1
1987	93.5	67.0	99.9	88.8	94.5	78.6	79.8
1988	93.9	67.2	101.8	88.9	95.8	77.6	78.8
Annual Growth Rate	-0.57%	-3.55%	0.16%	-1.06%	-0.39%	-2.28%	-2.14%
				Ray Scale Economies (Averaged across size classes)			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
1977	.947	.936	1.035	1.017	1.025	1.054	1.022
1978	.962	.954	1.034	1.032	1.033	1.051	1.047
1979	.968	.963	1.023	1.047	1.038	1.052	1.058
1980	.979	.976	1.017	1.046	1.036	1.061	1.037
1981	.981	.978	1.028	1.049	1.033	1.079	1.056
1982	.961	.957	1.026	1.044	1.025	1.049	1.068
1983	.944	.938	1.019	1.033	1.018	1.052	1.053
1984	.948	.944	1.035	1.032	1.018	1.051	1.044
1985	.941	.936	1.024	1.026	1.016	1.032	1.041
1986	.939	.935	1.012	1.021	1.017	1.011	1.012
1987	.937	.932	1.000	1.020	1.017	1.011	1.013
1988	.937	.930	1.015	1.020	1.024	1.014	1.019
Average	.955	.950	1.022	1.032	1.025	1.043	1.039

Table 5

Ray Scale Economies by Bank Size Class for the Preferred Models of the Stochastic Econometric and Thick Frontier Approaches by Year

Stochastic Econometric Approach Ray Scale Economies [MLE (by year)]

Bank Asset Size Classes
(M = million; B = billion)

	100M-200M	200M-300M	300M-500M	500M-1B	1B-2B	2B-5B	5B-10B	>10B	Overall
1977	1.005	1.006	1.016	1.023	1.035	1.045	1.061	1.089	1.035
1978	1.026	1.023	1.028	1.026	1.029	1.032	1.047	1.061	1.034
1979	1.033	1.031	1.024	1.019	1.014	1.010	1.023	1.026	1.023
1980	1.026	1.022	1.014	1.007	1.005	1.001	1.026	1.036	1.017
1981	1.044	1.039	1.034	1.022	1.017	1.009	1.021	1.035	1.028
1982	1.044	1.038	1.035	1.021	1.017	1.008	1.012	1.029	1.026
1983	1.031	1.026	1.024	1.011	1.011	1.004	1.012	1.031	1.019
1984	1.029	1.023	1.027	1.019	1.023	1.026	1.042	1.088	1.035
1985	1.031	1.024	1.022	1.014	1.011	1.012	1.024	1.057	1.024
1986	1.002	1.003	1.004	1.005	1.006	1.008	1.019	1.051	1.012
1987	.984	.988	.992	.993	.998	1.003	1.010	1.032	1.000
1988	.990	.985	1.002	1.011	1.018	1.024	1.032	1.055	1.015
Overall	1.020	1.017	1.018	1.014	1.015	1.011	1.027	1.049	1.022

Thick Frontier Approach Ray Scale Economies [Varying Cost Function, Stable Quartiles]

	100M-200M	200M-300M	300M-500M	500M-1B	1B-2B	2B-5B	5B-10B	>10B	Overall
1977	1.056	1.048	1.065	1.052	1.053	1.052	1.049	1.054	1.054
1978	1.101	1.087	1.092	1.060	1.045	1.026	.988	1.009	1.051
1979	1.141	1.124	1.110	1.066	1.035	1.012	.958	.967	1.052
1980	1.168	1.152	1.132	1.071	1.031	1.000	.953	.977	1.061
1981	1.150	1.133	1.120	1.084	1.071	1.053	.994	1.029	1.079
1982	1.097	1.081	1.070	1.054	1.049	1.044	.995	.999	1.049
1983	1.080	1.074	1.065	1.050	1.042	1.045	1.020	1.041	1.052
1984	1.073	1.069	1.058	1.050	1.050	1.049	1.012	1.048	1.051
1985	1.041	1.043	1.029	1.039	1.048	1.038	1.011	1.005	1.032
1986	1.031	1.025	1.015	1.018	1.018	1.008	.999	.993	1.011
1987	1.024	1.018	1.012	1.013	1.012	1.009	.997	1.011	1.011
1988	1.031	1.027	1.020	1.019	1.012	1.010	.993	.996	1.014
Overall	1.083	1.073	1.066	1.048	1.039	1.029	.997	1.008	1.043

APPENDIX

Means and standard deviations of the variables used in the equation system (1)-(2) are shown in Table A1 for the year 1988.[1] All data are from the Consolidated Reports of Condition and Income (Call Reports) except as noted.[2] Because of major changes in these reports, the study was started in 1977 rather than earlier. We included only banks that were in continuous operation over the 12-year period and that had more than $100 million in assets.

Only banks in states that permitted branching (limited or statewide) during any year of the sample were included. As of 1988, there were only 4 unit banking states (Colorado, Illinois, Montana, and Wyoming), although all states now allow branching. Bank mergers were treated as the acquisition of new deposits, assets, and factor inputs by the larger of the institutions involved, and the dummy variable M was added to account for the potential cost effects of these 391 mergers.[3] These restrictions eliminated approximately 11,500 banks with about one-third of bank assets. Banks were placed in size classes consistent with their average size over the 12-year period.

All value data were converted to real 1988 dollars using the GNP deflator prior to estimation. The GNP deflator may be a good price index for bank outputs, since bank deposits and loans are used to purchase the entire array of society's goods and services.[4] On an aggregate basis, there was a good correspondence between our deflated output series and that of the Bureau of Labor Statistics (BLS, 1989), which is based on actual physical measurements of checks processed, deposit and withdrawal activity, number of new loans made, and trust accounts serviced. Over 1977-86, the BLS series on bank output rose 40.4 percent, while our aggregate, cost share weighted series for the panel data set increased 43.8 percent over the same period.

---

1. These means are simple averages, which differ from some of the weighted figures discussed in the text. For example, average interest costs are only 59.1 percent of assets (S_3 plus S_4). However, total interest costs are about 75 percent of total bank assets, since larger banks are more interest-cost intensive.

2. The flow figures are the annual totals from the December Call Report, while the stock figures are averages of the prior December and current June and December Calls. The averaging avoids bias from growth or decline over the year.

3. This follows the treatment of airline mergers in Sickles, Good, and Johnson (1986, p. 151).

4. No direct price index for bank output exists for the full 1977-88 period.

Table A1
Summary of Data
(All 683 panel banks, 1988)

Cost Variables in Model*		**Mean**	**Std. Dev.**
TC	Total cost (expressed as a percent of assets).**	8.4%	1.3%
S_1	Labor share of total cost (percent).	18.6%	4.5%
S_3	Deposit interest share of total cost (percent).	40.1%	11.8%
S_4	Purchased funds interest share of total cost (percent).	19.0%	12.9%
Output Quantities and Input Prices			
Y_1	Demand deposits (as a percent of assets).**	16.7%	5.5%
Y_2	Retail (small) time and savings deposits (as a percent of assets).**	52.9%	14.5%
Y_3	Real estate loans (as a percent of assets).**	24.0%	10.2%
Y_4	Commercial and industrial loans (as a percent of assets).**	20.8%	9.1%
Y_5	Installment loans (as a percent of assets).**	13.4%	8.1%
P_1	Price of labor, $000 per year.	27.1	7.4
P_2	Price of physical capital (assumed to be proportionate to the replacement cost of a square foot of office space; taken from F.W. Dodge).	84.3	11.4
P_3	Interest rate on deposits.	4.8%	1.4%
P_4	Interest rate on purchased funds.	6.5%	1.2%

* The physical capital cost share (S_2) is excluded from the model to avoid perfect collinearity.

**Numbers are expressed relative to assets for exposition only. Regressions are based on raw data in $000.

All value figures are in constant 1988 dollars.

Source: Call Reports, except as noted above.

Notes

Notes

Notes