The Agency Problem Revisited: A Structural Analysis of Managerial Productivity and CEO Compensation in Large U.S. Commercial Banks

Shasha Liu [*]

Robin Sickles[†]

Abstract

The paper analyzes performance, incentives, and the inefficiencies that may arise due to agency problems and market power using a newly developed panel of large U.S. commercial banks that have too-big-to-fail nature. We use a structural model to characterize managerial efficiency, which complements technical efficiency in standard stochastic frontier models. We incorporate managerial decisions, bank-specific characteristics, and market competition in deriving managerial efficiency. Data on the 50 largest commercial banks in the U.S. during 2000 and 2017 are collected from the Call Reports, and are matched with CEO compensation from S&P's Execucomp database. The paper connects empirical evidence with economic theory and contributes to the literature on efficiency and management. The ultimate goal is to better understand the linkages among managerial performance, CEO compensation, and the size and scope of bank operations. Current results point to robust empirical findings. Economies of scale have steadily declined throughout the period, and are not positively related to managerial performance and CEO compensation. The size of a bank does not seem to be justified by the evidence in that larger banks offer larger bonuses and tend to have lower managerial efficiency and diminishing scale economies.

Keywords: Banking, Panel Data, Stochastic Frontier, Sources of Efficiency, Managerial Compensation

JEL Classification: C13, C33, D22, G21

[*] Quantitative Analytics Senior, Freddie Mac, 1551 Park Run Dr, McLean, VA 22102, USA

[†] Reginald Henry Hargrove Professor of Economics, Rice University, 6100 Main St, Houston, TX 77005, USA

1 Introduction

The banking industry has experienced major regulation changes in the last three decades. A period of banking deregulation removed restrictions on branching across states under the Riegle-Neal Interstate Banking and Branching Efficiency Act of 1994. Later, the Federal Reserve Board allowed commercial banks to underwrite insurance and securities through affiliates under the Financial Modernization Act (also known as Gramm-Leach-Bliley Act) of 1999. The number of banks declined by more than 50% after two decades of deregulation. Along with technological developments and financial innovations, banks have become much larger and provide more complex profiles of financial services. One direct consequence of banking deregulation has been financial consolidations which result in larger banks and more integrated banking system. Based on the Federal Reserve Statistics in 2017, the asset ratio of the 4 largest banks in the U.S., each with over $1 trillion in consolidated assets by 2017, has increased from 23% in 2000 to 45% in 2017, as shown in Figure 1. The asset share almost doubled during the period. The top 50 banks, each with over $10 billion in assets, have the asset ratio of 80% on average. This highly concentrated U.S. banking industry, along with the periods of regulation changes and technological developments, provide strong incentives for researchers to conduct studies to analyze bank performance,particularly the performance of too-big-to-fail banks.

A common aspect of large bank performance researchers look into is scale economies. Scale economies is one plausible reason for the increase in bank size. Larger banks may experience higher economies of scale, since larger scale of operations can have better diversification which may reduce marginal cost due to liquidity risk and credit risk. On the other hand, however, increased size can also trigger more risk-taking decisions and result in higher costs. Efficiency is also an important aspect to consider when analyzing large banks. Bank outputs can heavily depend on the technology which larger banks may be better equipped with. Thus, researchers need to look into efficiency to weigh the benefits and costs of increased bank size, as scale economies may only partially capture the effect of the increased concentration. We conduct an analysis on both scale economies and efficiency to have a better understanding of large banks. Specifically, we estimate scale economies and efficiency using a cost function approach which incorporates firm fixed effects derived from a structural framework to account for market competition and firm behaviors. Thus, the paper contributes to the literature on scale economies and efficiency with a structural approach to complement the commonly used reduced form method.

More importantly, the structural model introduced in our paper aims to identify a specific type of efficiency due to managerial decisions. Efficiency in standard stochastic frontier models (SFMs) results from a combination of, but not limited to, technical, regulatory, and managerial factors in the production process. Some factors are controlled by managers who may not have aligned interests with firm owners. The separation of management and ownership gives rise to a conflict of interest if the incentive mechanism fails, leading to the agency problem. The decomposition of the efficiency provides important insights on the efficiency that is fully controllable by management. We refer to such efficiency as managerial efficiency.

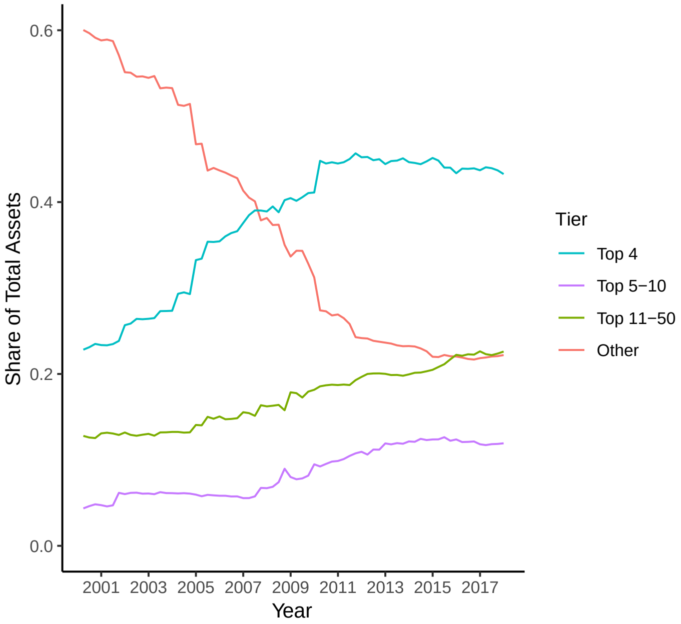

Figure 1: Shares of total assets

Therefore, the paper also contributes to the stochastic frontier literature by decomposing efficiency into managerial and technical efficiency, the latter of which in this context refers to the efficiency from uncontrollable factors.

From an econometric perspective, modeling managerial efficiency accounts for potential endogeneity in the SFMs. The paper uses a cost function to model managerial efficiency which is derived as a function of the cost frontier variables including all the outputs and input prices. In a standard SFM, the explanatory variables, inefficiency, and noise are assumed to be independent from each other. In such a setting, the cost frontier model would have a classic endogeneity problem due to omitted variables if it does not differentiate managerial inefficiency as a separate term or regards it as noise when managerial inefficiency is in fact correlated with output and impacts the cost of production. Thus, the parameter and efficiency estimates from a standard cost frontier model would be inconsistent. The paper deals with potential endogeneity in a cost function SFM to the extent that managerial inefficiency is correlated with outputs and impacts cost of production.

Another benefit of using the structural model is that it incorporates heterogeneous factors to explain inefficiency. The model accounts for managerial decisions, firm-specific characteristics, and market competition in deriving managerial inefficiency. By incorporating individual heterogeneity, the model has the advantage of identifying inefficiency due to environmental and heterogeneous factors, compared with a standard SFM which can only identify the inefficiency explained by environmental variables.

For the empirical analysis, we develop a dataset based on Call Reports for the top 50 banks during the period 2000-2017. The dataset provides comprehensive banking information that enables the use of a cost function, which requires data on total cost, outputs and input prices for each bank. Variations in bank characteristics exist even among the top banks. Thus, it is necessary to look into different tiers of banks for patterns and differences in performance over time. Since managerial performance is closely related to executive pay, the paper extends the analysis to examine the relationship between efficiency and CEO compensation (as additional materials in the Appendix). We use S&P's ExecuComp database, which provides data on annual CEO pay and CEO demographics.

The paper is one of the many steps we are executing in the research to examine the linkages among banking concentration, bank performance, and CEO compensations by providing empirical support for future studies on the linkages. The paper shows that scale economies of the sample average have declined throughout the period. The largest banks experience diseconomies of scale prior to the sample period. Technical efficiency follows the GDP growth and decreases during the Financial Crisis and has picked up since 2015. Managerial efficiency has declined for the Top 10 and slightly increased for smaller banks. The declining managerial efficiency with similar trends among the largest banks appears to indicate that they share higher costs due to possibly onerous liquidity and capital requirements the government has imposed since the Crisis compared with smaller banks due to the Dodd-Frank Act. The substantial dis-economies and lower managerial inefficiency among the largest banks draw attention to managerial contracts and performances. The paper finds a negative relationship between managerial efficiency and compensations, particularly pay-performance incentives.

The paper is organized as follows. Section 2 provides a literature review on measurement of efficiency, the agency problem, and scale economies. Section 3 introduces the structural model and the derivation for managerial efficiency. Section 4 describes the data we use. Section 5 provides estimation models and econometric methods, followed by results and implications in section 6.

2 Literature Review

2.1 Measurement of Efficiency

The fundamental idea of efficiency is to measure firm's performance in terms of the extent to which outputs are produced given a certain amount of inputs. A number of approaches are available to measure efficiency. Studies on productivity and efficiency in the banking industry have extensively used two approaches, the non-parametric Data Envelopment Analysis (DEA) and the parametric Stochastic Frontier Analysis (SFA). The DEA approach is based on the works of Debreu (1951), Shephard (1953), and Farrell (1957). It is first used to estimate production efficiency by Charnes, Cooper, and Rhodes (1978). The DEA applies a linear programming technique to search for the best-practice production units to form the efficient frontier, which is a piece-wise linear convex hull of the most efficient units. The SFA approach, introduced by Aigner, Lovell, and Schmidt (1977) and Meeusen and van den Broeck (1977), estimates a parameterized efficiency frontier with an error term composed of a one-sided (non-negative) inefficiency and a two-sided noise term.

For the purpose of this paper, the SFA is the preferred approach over the DEA as it is more easily implemented in standard regression models, albeit with adjustments made to deal with potential endogeneity of the efficiency with input choices.[1] A major drawback of the DEA is that in its standard form as in Charnes et al. (1978) it fails to account for statistical noise that is usually present in real data. Although relatively recent developments by Simar and Wilson (2013) have addressed in part these shortcomings, the very nature of the optimization problem that DEA uses to solve for efficiency does not have a natural complement in the economists' regression models, especially in regard to the structural models we develop below. Thus, the important issues we are trying to analyze, such as insights into market structure and firm behavior cannot be analyzed using models familiar to most economists. The SFA, on the other hand, specifies inefficiency as a (one-sided) latent variable in the regression model.[2] For these reasons we choose the SFA approach in our structural modeling efforts to identify scale economies and to account for market competition.

The general form of a SFM is:

$$y_{it} = \alpha + f(X_{it};\beta) + v_{it} \pm u_{it},\ i = 1, ..., N;\ t = 1, ..., T, \tag{1}$$

where y_{it}, depending on the function used for the SFM, can be total output or cost of a production unit, X_{it} is a vector of explanatory variables (inputs for a single output production function or prices and output for the corresponding cost function dual), $f(X_{it};\beta)$ is a frontier function (e.g. Cobb-Douglas or translog), is the frontier intercept, v_{it} is the usual two-sided noise term that is assumed to be independently and identically distributed, and $u_{it} \geq 0$ is the inefficiency of which the sign depends on the function for the SFM, e.g - for a production function and + for a cost function. The general form shows that $u = 0$ when the unit is fully efficient, and larger (smaller) u means that the unit is more output (cost) efficient. One can think of SFA as a composed error model in which one component is statistical noise and the other part is one-sided inefficiency.

Earlier studies by Pitt and Lee (1981) and Schmidt and Sickles (1984) on stochastic frontier models using panel data assume time-invariant inefficiency, an assumption that might not be valid for longer panels. Later, Cornwell, Schmidt, and Sickles (1990) extend the panel model in (4.1) by allowing for heterogeneity

$$\alpha_{it} = \delta_i W_t = \delta_{i1} + \delta_{i2} t + \delta_{i3} t^2, \tag{2}$$

where $\delta_{i1}, \delta_{i2}, \delta_{i3}$ are firm specific and t is the time trend.

Based on a similar idea, Battese and Coelli (1992) propose a time decay model which specifies u_{it} as $g(t)u_i$, and $g(t)$ is defined as

$$g(t) = \exp(-\sigma(t - T_i), \tag{3}$$

[1] See, for example, Olley and Pakes (1996); Schmidt and Sickles (1984); Kutlu (2018).

[2] For a more detailed comparison of DEA and SFA applied in the banking industry, see Ferrier and Lovell (1990) and Bauer et al. (1998).

where T_i is the last period of ith panel, and σ is the decay parameter. Models of Cornwell et al. (1990) and Battese and Coelli (1992) are the primary models used in empirical work to account for time varying inefficiency.

However, Greene (2005) points out that time-varying inefficiency can contain firm-specific heterogeneity and thus uses fixed effects and random effects models to account for heterogeneity which can be correlated with the frontier variables. The true fixed effects model (TFE), as Greene has labeled it, has the general form

$$\begin{aligned} y_{it} &= \alpha_i + \boldsymbol{\beta}'\boldsymbol{x}_{it} + v_{it} \pm u_{it} \\ v_{it} &\overset{iid}{\sim} N(0, \sigma_v^2) \\ u_{it} &\overset{iid}{\sim} F(\sigma_u^2), \quad i = 1, ..., n, \; t = 1, ..., T, \end{aligned} \tag{4}$$

where α_i is the fixed effects, the error term v_{it} and one-sided non-negative inefficiency u_{it} are assumed to be independently distributed of each other. One can choose a particular distribution of u_{it}, e.g. half-normal or truncated normal, from the family of distributions denoted as F with a scale parameter σ_u.

This specification of the TFE model defines inefficiency as a shortfall in production due to factors that are not time invariant and correlated with the regressors. We use it in the structural model in order to control for such time invariant heterogeneity factors. These factors would include firm characteristics that impact the manager's utility from shirking, such as corporate culture and monitoring environment that do not change frequently for relatively long periods of time.

2.2 Efficiency Due to the Agency Problem and Market Power

A number of studies have motivated researchers to model efficiency losses related to the agency problem. Williamson (1963) incorporated managerial discretion in firms which separate management from ownership. Based on his model, the managers discretionarily maximize personal utilities, which are not necessarily aligned with the utilities of the firm owners. Later, Leibenstein (1966) proposed X-Efficiency theory, which argues for an undefined inefficiency that is greater in magnitude than allocative inefficiency. A major source of X-Inefficiency is an agent's motivation,which leads to efficiency lost. This is the agency problem.

Previous studies also have shown support for estimating a different source of inefficiency inspired by the implications of agency theory. These studies attempt to explain a portion of the total inefficiency that can be identified with the agency problem using various models, some structural and some reduced form-based models in which proxies for the way in which the agency problem may manifest itself as productive or cost inefficiency are used as factors for the latent agency effect. The few studies that have used a structural approach to incorporate the agency problem in the SFA include those by Gagnepain and Ivaldi (2002), who estimate a structural model of efficiency based on agency theory using data of the French public transit system in which firms are regulated. However, they fail to account for market competition which is found to be linked with efficiency in the estimation. Kutlu and Wang (2018) incorporate market competition when estimating efficiency, but they do not distinguish the inefficiency due to the agency problem from the one-sided inefficiency in standard SFMs. Kutlu et al. (2018) construct a structural type of inefficiency aside from the non-structural inefficiency in SFMs. It is the closest work to our paper, but the Kutlu et al. (2018) model does not account for firm-specific factors in deriving the structural inefficiency. The paper modifies their approach by allowing firm heterogeneity in managerial decisions in order to introduce structural inefficiency, which is the managerial inefficiency in this context.

3 Structural Model

We next develop a structural model that makes explicit the role of agency-based incentive inefficiency in the provision of banking intermediation services, as well as other potential sources, the latter of which are introduced as a latent unobservable inefficiency composite term.

We begin with a somewhat stylized model in which there is a single manager in each firm that produces one homogeneous output in an oligopolistic market. Firms compete by choosing how much to produce and the price is determined by the total output in the market. The oligopoly assumption for the banking industry is consistent with the current organization structure of the banking industry and the fact that the largest four banks in the U.S. (JPMorgan Chase, WellsFargo, Bank of America, Citigroup) take up 46 % of total consolidated assets by 2017. The manager of a firm is assumed to be the only decision maker. Owner/shareholders only observe the profits resulting from the manager's decision

$$\pi_i(\boldsymbol{q}, s_i) = P(Q)q_i - C_i(q_i, s_i), \tag{5}$$

where q_i is the quantity of output of firm i, $q = (q_1, q_2, ..., q_n)$ is a vector of outputs from each of n firms, $Q = \sum_{i=1}^{n} q_i$, $s_i \geq 1$ is the level of shirking manager i chooses and is determined by the manager's utility maximization problem in the next section. Note that a firm's profit also depends on the output decisions of other firms' managers.

The cost function with unobserved managerial effort s_i, denoted as C_i, is assumed to be separable such that

$$C_i(q_i, s_i) = C_i^F(q_i)s_i, \tag{6}$$

where $s_i \geq 1$ and represents the cost associated with the optimal level of shirking chosen by the manager to maximize her utility. Therefore, the more the managerial slack, the higher the cost incurred to a firm. C_i^F is the frontier (minimum/efficient) level of costs. We generalize this and allow for the frontier to be impacted by random shocks as well as other sources of cost inefficiency when we specify the stochastic version of equation (6).

The cost function is a convenient vehicle for analyzing the agency problem when firms are producing multiple exogenous outputs and are price takers in the input markets. The large banks, however, can exercise their market power to select outputs, and thus outputs are potentially endogenous. We account for endogeneity in the outputs in the cost function estimation using instrumental variables. Incorporating managerial effort into the standard cost frontier model, the cost function becomes

$$\ln C_i = \ln C_i^F + \ln s_i + u_i + v_i, \tag{7}$$

where u_i is one-sided inefficiency and v_i is the error term as in equation (1) above.

As the manager maximizes her own utility rather than the firm's profit, the objective of the manager of firm i is

$$\max_{q_i, s_i} \; U_i(\boldsymbol{q}, s_i) = \alpha_i \pi_i(\boldsymbol{q}, s_i) + \tilde{R}_i(s_i) + \tilde{B}_i, \tag{8}$$

where $\boldsymbol{q} = (q_1, ..., q_N)$, $\alpha_i \in (0,1)$ is the percent of profit (e.g. pay-performance bonuses) paid to the manager, $\tilde{R}_i(s_i)$ is the self-rewarding utility that is firm-specific, and $\tilde{B}_i$ is the baseline salary of the manager. We assume that $\tilde{R}_i(s_i) \geq 0$, $\tilde{R}_i'(s_i) > 0$, $\tilde{R}_i''(s_i) < 0$ so that the self-rewarding utility increases in s_i at a decreasing rate. The utility function can be normalized without loss of generality and becomes

$$U_i(\boldsymbol{q}, s_i) = \pi_i(\boldsymbol{q}, s_i) + R_i(s_i). \tag{9}$$

The decision-making process is assumed to be a simultaneous game in which the manager chooses the effort level and output quantity at the same time. The first order conditions of the maximization problem yield the following equations:

$$\begin{aligned} P(Q) &= -P'(Q)q_i + C_i^{F'}(q_i)s_i \\ s_i &= R_i'^{-1}(C_i^F). \end{aligned} \tag{10}$$

The econometric models can then be derived as

$$\begin{aligned} P_i &= -P_i'(Q)q_i + C_i^{F'}(q_i)R_i'^{-1}(C_i^F) + \epsilon_i \\ \ln C_i &= \ln C_i^F(q_i) + \ln R_i'^{-1}(C_i^F) + u_i + v_i, \end{aligned} \tag{11}$$

where $\ln R_i'^{-1}(C_i^F) \geq 0$ captures the inefficiency from managerial slack, $u_i \geq 0$ accounts for technical and allocative inefficiency as in the stochastic frontier literature, and ϵ_i and v_i are the usual idiosyncratic noise terms. Input allocations are assumed to be optimal such that banks only have technical inefficiency, an assumption supported by findings in Inanoglu et al. (2016) and Al-Sharkas, Hassan and Lawrence (2008). Output prices are bank-specific since bank prices are not regulated, and banks can charge different interest rates on their products depending on their local markets.

Since banks produce more than one output, the model is extended to a multi-output scenario. The first order conditions of profit maximization of firm i that produces M types of outputs are

$$\sum_{h=1}^{M} \frac{\partial P_h(\boldsymbol{Q})}{\partial Q_m} q_{hi} + P_m(\boldsymbol{Q}) - \frac{\partial C_i^F(\boldsymbol{q}_i)}{\partial q_{mi}} s_i = 0, \;\; m = 1, ..., M, \tag{12}$$

where the output quantities produced by firm i is $\boldsymbol{q}_i = (q_{1i}, ..., q_{Mi})$, and total quantities in the market is $\boldsymbol{Q} = (Q_1, ..., Q_M)$ with demands represented by $(P_1(\boldsymbol{Q}), ..., P_M(\boldsymbol{Q}))$. Allowing firm-specific prices by introducing the error terms, the econometric model for each output becomes

$$P_{mi} = -\sum_{h=1}^{M} \frac{\partial P_h(\boldsymbol{Q})}{\partial Q_m} q_{hi} + \frac{\partial C_i^F(\boldsymbol{q}_i)}{\partial q_{mi}} s_i + \epsilon_{mi},\ m = 1, ..., M. \tag{13}$$

4 Data

We collect banking data from quarterly consolidated reports of balance sheets and income statements, also known as the Call Reports. All national banks, state member banks, and insured state non-member banks are required to file on a quarterly basis on the last calendar day of each quarter. The bank data consist of detailed information on a bank's various assets, liabilities, capital structure, expenses, and geographical characteristics. The Call Reports are expressed on a pro-forma basis, which accounts for mergers, controls for survival bias, and does not distort the measurement of banks' growth. The data methodology is useful among banks to calibrate credit risk models.[3] We focus on commercial banks and use total assets as a measure of bank size. Total assets are less prone to changes in the internal models than risk-weighted assets. The availability and straightforward definition of total assets also make it a favorable indicator among central bankers and researchers.

There is no complete consensus on the exact guideline for the choice of what constitutes banks' inputs and outputs. Three methods have been used to define inputs and outputs in the banking industry: the intermediation approach (or the asset approach), the user cost approach, and the value-added approach. The major difference among these approaches is the way deposits are defined.[4] Deposits can be inputs or outputs in the user cost approach, but they are usually considered as outputs in the value-added approach. However, both the user cost and value-added approaches encounter measurement difficulties and require detailed data on transactions hard to obtain. The asset approach is more consistent with the banking data (Adams et al., 1999) and more popular among various studies of banking efficiency (e.g. Hughes, Mester and Moon, 2001; Drake and Hall, 2003; Weill, 2004; Feng and Serletis, 2010; Davies and Tracey, 2014; Inanoglu, Jacobs Jr., Liu, and Sickles, 2016). Therefore, we adopt the asset approach to define inputs and outputs. Specifically, we choose number of employees, premises & fixed assets, and interest-bearing deposits as inputs; real estate loans, commercial & industrial loans, consumer loans, and securities as outputs. Under

[3] For details, see Inanoglu and Jacobs (2009).

[4] The asset approach assumes that banks are intermediaries whose main function is to collect deposits from savers and transform them into loans and financial investments. The user cost approach considers a financial instrument as an output only when the net revenue exceeds the opportunity cost of funds or the costs of liability are smaller than the opportunity cost. Otherwise, it is an input. The value-added approach, however, does not exclusively differentiate inputs from outputs. It determines whether financial products are outputs, inputs, or intermediates depending on how much value the categories of the products generate. For a more detailed discussion of the approaches, see Berger and Humphrey (1992).

the asset approach, the measure of bank outputs is the dollar value of loans and securities, and the measure of total cost is interest costs plus operational costs.

After sorting the banks based on the value of total assets as of the last quarter of 2017, we choose the top 50 commercial banks established in the U.S. and study the period from 2000 to 2017. The year 2000 is a proper starting point to study how periods of banking deregulation change the bank efficiency because restrictions on entry and products have been removed by 1999 and some outputs are measured in a different category prior to 2000. Table 1 lists the names of the top 50 banks ranked by total assets by the last quarter of 2017. All these banks have at least $10 billion in total assets. Table 2 provides descriptive statistics for the banks. Note that a wide variation exists in the output variables and total assets among the top 50 banks. Input prices have a relatively smaller variance, although the difference among outliers can still be drastic. This indicates that it is necessary to look into different tiers of banks for a proper comparison.

Table 1: Top 50 U.S. commercial banks

Rank	Name	Total assets	Total equity capital	Equity ratio	Total deposits
1	JPMorgan Chase Bank NA	1.88e+09	1.86e+08	9.89%	7.67e+08
2	Bank of America NA	1.54e+09	1.82e+08	11.81%	7.35e+08
3	Wells Fargo Bank NA	1.53e+09	1.46e+08	9.54%	7.33e+08
4	Citibank NA	1.21e+09	1.25e+08	10.33%	3.14e+08
5	U.S. Bank NA	4.00e+08	4.17e+07	10.43%	2.10e+08
6	PNC Bank NA	3.25e+08	3.42e+07	10.52%	1.60e+08
7	Bank of New York Mellon	2.61e+08	2.37e+07	9.08%	4.41e+07
8	Capital One NA	2.55e+08	3.31e+07	12.98%	1.73e+08
9	Branch Banking & Trust Company	1.90e+08	2.44e+07	12.84%	9.66e+07
10	SunTrust Bank	1.77e+08	2.15e+07	12.15%	1.08e+08
11	HSBC Bank NA	1.58e+08	2.04e+07	12.91%	8.66e+07
12	Fifth Third Bank	1.23e+08	1.48e+07	12.03%	6.23e+07
13	The Northern Trust Company	1.21e+08	8.09e+06	6.69%	1.68e+07
14	KeyBank NA	1.19e+08	1.33e+07	11.18%	6.49e+07
15	Regions Bank	1.08e+08	1.41e+07	13.06%	5.33e+07
16	MUFG Union Bank NA	1.04e+08	1.44e+07	13.94%	4.64e+07
16	Manufacturers and Traders Trust Company	1.04e+08	1.26e+07	12.12%	5.25e+07
18	BMO Harris Bank NA	9.59e+07	1.35e+07	14.08%	4.47e+07
19	Huntington NB	9.13e+07	9.92e+06	11.28%	4.87e+07
20	Bank of the West	7.87e+07	1.06e+07	13.60%	4.59e+07
21	Compass Bank	7.59e+07	1.06e+07	13.97%	4.19e+07
22	Comerica Bank	6.28e+07	6.50e+06	10.35%	2.33e+07
23	Zion Bank NA	5.80e+07	6.68e+06	11.52%	2.55e+07
24	Silicon Valley Bank	4.42e+07	3.30e+06	7.47%	4.71e+06
25	City NB	4.20e+07	3.44e+06	8.18%	1.46e+07
26	First Tennessee Bank NA	3.61e+07	4.31e+06	11.94%	1.98e+07
27	East West Bank	3.26e+07	3.36e+06	10.31%	1.69e+07
28	Banco Popular de Puerto Rico	3.05e+07	3.27e+06	10.71%	1.87e+07
29	First-Citizens Bank & Trust Company	3.01e+07	2.81e+06	9.34%	1.58e+07
30	BOKF, NA	2.83e+07	2.86e+06	10.12%	1.79e+07
31	Frost Bank	2.79e+07	2.87e+06	10.27%	1.38e+07
32	First National Bank of Pennsylvania	2.74e+07	3.92e+06	14.58%	1.46e+07
33	Synovus Bank	2.73e+07	2.83e+06	10.38%	1.63e+07
34	Associated Bank NA	2.67e+07	2.73e+06	10.88%	1.53e+07
35	Iberiabank	2.44e+07	3.16e+06	12.97%	1.34e+07
36	Whitney Bank	2.39e+07	2.59e+06	10.84%	1.22e+07
37	Umpqua Bank	2.26e+07	3.77e+06	16.69%	1.18e+07
38	CIBC Bank USA	2.26e+07	4.72e+06	20.86%	1.19e+07
39	Texas Capital Bank NA	2.20e+07	1.90e+06	8.65%	9.92e+06
40	Pacific Western Bank	2.19e+07	4.55e+06	20.79%	9.09e+06
41	Commerce Bank	2.17e+07	2.11e+06	9.74%	1.16e+07
42	Valley NB	2.10e+07	2.31e+06	11.02%	1.13e+07
43	BNY Mellon, NA	2.10e+07	3.17e+06	15.08%	1.46e+07
44	TCF NB	2.02e+07	2.25e+06	11.21%	1.29e+07
45	Prosperity Bank	1.98e+07	3.34e+06	16.87%	1.07e+07
46	UMB Bank, NA	1.89e+07	1.59e+06	8.43%	9.81e+06
47	Bank of the Ozarks	1.87e+07	3.04e+06	16.25%	1.27e+07
48	First Hawaiian Bank	1.80e+07	2.21e+06	12.29%	9.47e+06
49	First National Bank of Omaha	1.79e+07	1.71e+06	9.57%	9.98e+06
50	MB Financial Bank, National Association	1.76e+07	2.60e+06	14.76%	7.70e+06

The statistics are presented as of 2017Q4. Dollar figures are in thousand $.

Table 2: Descriptive statistics for the top 50 banks

	Mean	**Std. Dev.**	**Min.**	**Max.**
Total Cost (C)	1.72e+06	4.11e+06	3.03e+03	3.44e+07
Output Variable				
Real Estate Loans (RE)	3.27e+07	7.18e+07	3.30e+04	4.68e+08
Commercial & Industrial Loans (CI)	1.59e+07	3.29e+07	2.71e+04	2.36e+08
Consumer Loans (CON)	9.85e+06	2.50e+07	4.82e+03	1.70e+08
Securities (SEC)	2.36e+07	5.70e+07	2.06e+04	3.71e+08
Input Price Variable				
Labor Price (w_L)	50.39	29.39	1.84	209.08
Capital Price (w_K)	25.33%	0.23%	0.50%	322.64%
Deposit Price (w_D)	0.94%	0.01%	0.01%	16.51%
Output Price Variable (%)				
Price of Real Estate Loans (p_1)	0.51	0.01	4.31e-05	14.11
Price of Commercial & Industrial Loans (p_2)	0.75	0.01	4.33e-05	13.68
Price of Consumer Loans (p_3)	2.16	0.04	1.58e-04	43.81
Price of Securities (p_4)	2.31	0.01	0.03	9.26
Control Variable				
Net Charge-offs to Loans	0.62	0.91	-2.13	16.79
Loss Allowance to Loans	1.56	0.74	7.92e-03	6.69
Age of bank	111	52.47	18	213
Total Assets	1.31e+08	3.12e+08	2.91e+05	1.91e+09

Note: The table shows summary statistics of the variables used in the translog cost function and the system of equations over the period 2000Q1-2017Q4. Total number of observations is 3600. Non-price dollar figures and labor price are in thousand $.

For data on CEO compensations we use Standard and Poor's Execucomp database, which provides time series data on executive compensation collected from a company's annual proxy since 1992. Even though reporting regimes have changed over the period, the main compensation variables, which are used in the analysis, are continuously reported with consistency. Multiple executives for a given fiscal year of a bank are reported, and most banks report details on 5 executives. Data on salary, bonus, stock and options awards, non-equity incentive plans, pensions and other compensations are collected. The compensations are averaged across executives to represent a CEO compensation package a bank offers in a given year, and are matched with banking data by ticker symbol.

5 Estimation

5.1 Derivation of System of Equations

To estimate the models, we need to specify functional forms for the market demand and the self-rewarding utility. In an oligopolistic market, banks produce according to the following inverse market demand functions

$$\ln P_{im} = \theta_{0m} - \sum_{h=1}^{4} \theta_{hm} \ln Q_h + \theta_{5m} X_i + e_{im}, \ i = 1, ..., 50, \ m = 1, ..., 4, \tag{14}$$

where P_{im} is the market price of output m, Q_h is the total quantity of output h, X_i are control variables, and e_{im} is the error term.[5] Note that unobserved factors that affect P_{im} can be correlated with Q_h. An expected decrease in the interest rate by the central bank, for example, increases the demand for certain loans since the cost of capital investment is expected to be lower, resulting in higher loan prices. In order to deal with this classic endogeneity problem due to omitted variables and to the possibility that managers exercise substantial discretion on the choice of the outputs, we instrument the outputs in equation (15) using lagged outputs.[6] Generalized Method of Moments estimation (GMM) is used to estimate the system of demand equations along with the system of supply equations.

The self-rewarding utility function of managers i takes the form

$$R_i(s_i) = r_i s_i^{\tau}, \tag{15}$$

where $r_i > 0$ and $\tau \in (0, 1)$, so that $R_i(s_i) > 0$, $R_i'(s_i) > 0$, $R_i''(s_i) < 0$. The utility function assumes that a manager's utility from shirking also depends on exogenous firm-specific characteristics, as shown by r_i, such as the board's power, monitoring capacity, and competitive culture within the firm that do not change in a relatively short period of time. If powerful executives serve on the board, for example, collusion instead of monitoring among the executives is more likely to take place, which leads to higher utility from shirking. On the other hand, if a firm has a competitive culture which incentivizes its employees to monitor the manager and the board, utility from shirking decreases because of higher opportunity cost.

The inverse of $R_i'(s_i)$ becomes

$$\begin{aligned} R_i'^{-1}(C_i^F) &= (r_i\tau)^{1/(1-\tau)} C_i^{F^{1/(\tau-1)}} \\ &= \gamma_i C_i^{F^{\delta}}, \end{aligned} \tag{16}$$

where $\gamma_i > 0$ and $\delta < -1$. This inverse function is the key to differentiate the structural inefficiency from the non-structural technical inefficiency in the stochastic frontier literature.

With our panel data, the system of equations for firm i in time t is

$$\begin{aligned} P_{mit} &= \sum_{h=1}^{4} \theta_{hm} P_{hit} \frac{q_{hit}}{Q_{mt}} + \frac{\partial C_{it}^F(\boldsymbol{q}_{it})}{\partial q_{mit}} \gamma_i C_{it}^{F^{\delta}} + \epsilon_{mit},\ m = 1, ..., 4 \\ \ln C_{it} &= \ln \gamma_i + (1+\delta) \ln C_{it}^F + u_{it} + v_{it}. \end{aligned} \tag{17}$$

[5] The form of inverse demand functions is chosen in order to facilitate estimation and interpretation. We have examined other functional representations of the inverse demand equations, and our results are not qualitatively different from using semi-log specifications as well as models with second-order interactions.

[6] We test the validity of these instruments using Hansen's J-test of overidentifying restrictions. We fail to reject J-test's null hypothesis that instruments are uncorrelated with the error term, and the excluded instruments are correctly excluded from the equation.

5.2 Stochastic Frontier Models with True Fixed Effects

In the case of a cost function, the TFE model becomes

$$C_{it} = \alpha_i + \boldsymbol{\beta}' \boldsymbol{X}_{it} + u_{it} + v_{it}. \tag{18}$$

Greene (2005) uses a maximum likelihood dummy variable approach to estimate the fixed effects and the cost parameters. He shows that computation is feasible even with a large number of nuisance parameters. The derived cost model resembles the TFE stochastic frontier model with $\ln \gamma_i$ as firm fixed effects.[7]

Recall that managerial effort s_{it} is dependent on γ_i, the unobserved frontier production cost function C_{it}^F, and a constant δ:

$$\begin{aligned} s_{it} &= \gamma_i C_{it}^{F^{\delta}} \\ \ln s_{it} &= \ln \gamma_i + \delta \ln C_{it}^F. \end{aligned} \tag{19}$$

In this setup, fixed effects partially account for managerial efforts. The unobserved cost, the elasticity of which is magnified by the constant ($|\delta| > 1$), plays a more informative role than firm effects in describing managerial efforts. This is consistent with the finding in Triebs and Kumbhakar (2018), which shows that management quality is only explained by a small percent of the variances of fixed effects in the standard frontier models.

5.3 Translog Cost Function

To model the unobserved cost, we use the translog cost function (Christensen, Jorgensen and Lau, 1973; Diewert, 1974), which is a popular flexible functional form using a second order logarithmic Taylor series expansion. It is useful for deriving various measurements with policy implications, such as input price elasticities, scale economies, and scope economies. For the multi-output cost function, the translog function using the simplified notation for panel data becomes

$$\begin{aligned} \ln C = \alpha_0 + \sum_{i=1}^{4} \alpha_i \ln q_i + \sum_{k=1}^{3} \beta_k \ln w_k + \frac{1}{2} \sum_{i=1}^{4} \sum_{j=1}^{4} \alpha_{ij} \ln q_i \ln q_j + \\ \frac{1}{2} \sum_{k=1}^{3} \sum_{l=1}^{3} \beta_{kl} \ln w_k \ln w_l + \sum_{i=1}^{4} \sum_{k=1}^{3} \delta_{ik} \ln q_i \ln w_k + Z, \end{aligned} \tag{20}$$

where w_k are the input prices of labor (L), fixed capital (K), and deposits (D), q_i are real estate loans (RE), commercial & industrial loans (CI), consumer loans (CON), and securities (SEC), and Z are control variables. Since the function is continuously differentiable,

[7] We choose half-normal distribution for u_{it}. A test for heteroscedasticity indicate that different variabilities exist in u_{it}. We use OCC districts (Central, Northeastern, Southern, and Western) in which banks are located to account for heteroscedasticity in u_{it}.

parameters are symmetric, i.e. $\alpha_{ij} = \alpha_{ji}$ and $\beta_{kl} = \beta_{lk}$. Homogeneity in input prices requires $\sum_{k=1}^{3} \beta_k = 1$, $\sum_{l=1}^{3} \beta_{kl} = \sum_{k=1}^{3} \beta_{lk} = \sum_{k=1}^{3} \delta_{ik} = 0$. The control variables include time trends, total assets, bank age, and loan quality to proxy credit risk. Net charge-offs ratio and loss allowance ratio are used to represent loan quality. A higher percentage of charge-offs or loss allowance ratio reduces real outputs and increases expenditure to maintain risks and outputs. Thus, estimated scale economies would be biased if credit risk was not accounted for. To account for endogeneity in the outputs, we use instrumental variables. Lagged total output quantity Q_i for each output q_i in the market, output prices and inputs with their lags are chosen as the instrumental variables, and they are tested to be valid instruments.

The measure of economies of scale is defined as

$$Scale \equiv \frac{C}{\sum_i q_i C_i(q)}, \tag{21}$$

where $q = (q_1, ..., q_4)$ is the vector of output quantities, and $C_i(q) = \frac{\partial C(q)}{\partial q_i}$ is the marginal cost of a particular output. $Scale > 1$ when economies of scale exist. Using the translog function, a common expression for scale economies is

$$Scale \equiv \frac{1}{\sum_i \partial \ln C / \partial \ln q_i}, \tag{22}$$

which depends on the elasticity of cost with respect to outputs. Thus, economies of scale exist when the percent increase in cost is less than that in the output quantity, which leads to a slower increase in the average cost. Note that both positive and negative scale economies can exist given the translog function. For the translog cost function, the economies of scale are derived as follows

$$\sum_{i=1}^{4} \frac{\partial \ln C}{\partial \ln q_i} = \sum_{i=1}^{4} \alpha_i + \sum_{i=1}^{4} \sum_{j=1}^{4} \alpha_{ij} \ln q_j + \sum_{i=1}^{4} \sum_{k=1}^{3} \delta_{ik} \ln w_k. \tag{23}$$

Economies of scope provide another aspect on cost reduction among output pairs. Economies of scope exist when producing multiple outputs together by a firm reduces the cost of producing the outputs separately by different firms (Baumol et al., 1982). A simplified expression for the economies of scope of producing two outputs is

$$Scope \equiv \frac{C(\bar{w}, \hat{q}_1, q_2^m) + C(\bar{w}, q_1^m, \hat{q}_2) - C(\bar{w}, \bar{q}_1, \bar{q}_2)}{C(\bar{w}, \bar{q}_1, \bar{q}_2)}, \tag{24}$$

where $\bar{w}$ and $\bar{q}_i$ are arbitrary values of input prices and output i, q_i^m is a small value of output, and $\hat{q}_i = \bar{q}_i - q_i^m$.[8] We choose the sample mean for $\bar{q}_i$ and $\bar{w}$ and the sample minimum for q_i^m. $Scope > 0$ when economies of scope exist. The expression can be easily extended to four outputs.

[8] Baumol et al., 1982 use zero value for q_i^m, but in the use of translog function $\ln q_i^m$ is undefined. Thus, we choose an arbitrarily small number.

We calculate the Allen-Uzawa and Morishima elasticities of substitution (AES and MES respectively) with the estimates from the translog function. Given the parameter estimates from the translog function, we can write the AES, θ_{lk}, and price elasticities of demand, η_{lk}, as follows

$$\theta_{lk} = \begin{cases} 1 + \beta_{lk}/s_l s_k, \ l \neq k \\ (\beta_{kk} + s_k^2 - s_k)/s_k^2, \ l = k \end{cases} \tag{25}$$

$$\eta_{lk} = \begin{cases} \frac{\beta_{lk} + s_l s_k}{s_l} = \theta_{lk} s_k, \ l \neq k \\ \frac{\beta_{kk} + s_k^2 - s_k}{s_k} = \theta_{kk} s_k, \ l = k, \end{cases} \tag{26}$$

where β_{lk} is the coefficient from the translog function, and s_l or s_k is the input price elasticity of cost or factor share. Own-price elasticity is expected to be negative. Market power can influence the elasticity, and thus a more inelastic demand for a normal good may imply concentrated market power.[9] The MES based on the definition from Blackorby and Russell (1989) are

$$M_{lk} = \eta_{kl} - \eta_{ll}. \tag{27}$$

Unlike the AES, the MES preserve the essential features of the original Hicksian concept. Thus, the MES have the advantage of being a curvature measure and providing information on relative factor shares. Specifically, the MES can assess the effects of changes in price or quantity ratios on relative factor shares, while the AES do not serve the purpose.

5.4 Estimation of the System of Equations

The estimation of the system of equations involves two steps. In the first step, we estimate the cost function and obtain estimators of the fixed effects and the cost parameters. To distinguish the estimates of the cost function from δ, we need to identify δ from the simultaneous non-linear equations in (17) as the second step.

We use GMM to estimate the simultaneous non-linear supply equations, along with the system of linear demand equations. In particular, iterative GMM is applied to the complete system of equations. The iterative estimator may have better finite-sample properties according to Hall (2005). When the model is correctly specified and each equation satisfies the rank condition, iterative GMM is more efficient in estimating the equations jointly than the 2SLS or IV, which estimate each equation in the system one at a time. The instrumental variables for the supply equations (17) are input prices and their lags, lagged total outputs, and firm effects estimated from the translog cost function. The instruments for the demand equations (14) are lagged total outputs.

Once we obtain the key estimates from the complete system of equations, we can calculate managerial inefficiency by

$$MIE_{it} \equiv \frac{s_{it} - s_t^*}{\sum_i (s_{it} - s_t^*)}, \tag{28}$$

[9] All the inputs, i.e. labor, fixed capital, and deposits, are considered as normal goods.

where $s_t^* = \min(s_{it})$ at time t. Managerial efficiency is then $ME_{it} \equiv 1 - MIE_{it}$, which measures a bank's ability to contain production cost relative to the bank with best practice.

6 Results

Estimates of the translog cost function using the true fix effects are presented in Table 3. The scale economies estimated at sample means is 0.9822. The magnitude of the linear time trend is small, indicating little to no technological progress during the sample period. The environmental variable, OCC District, is significant in explaining the heteroscedasticity in technical inefficiency. The magnitude of the variable indicates that the regulation district a bank belongs to accounts for a large percentage of variance in technical inefficiency. Table 4 shows economies of scope estimates for all output pairs. Large banks have economies of scope among all pairs, indicating that options are available to cross-subsidize multiple outputs. Producing real estate loans and investing in securities have the largest economies of scope, followed by producing commercial & industrial loans and consumer loans. The elasticities of substitution and elasticities of demand for the inputs are in Table 5. The top panel is symmetric AES estimates of elasticities. The middle panel for the asymmetric MES excludes the diagonal since it contains no information. The last panel shows the elasticities of demand. Capital has the largest elasticity, and labor and deposits have similar elasticities. Table 6 presents parameter estimates for the system of equations in the second step estimation. The magnitude of technical inefficiency shows that it is the major source of inefficiency that impedes bank performance. Even though the magnitude of managerial inefficiency is smaller, it is substantial and accounts for approximately 20% of bank-level inefficiency.

Table 3: Parameter estimates for the true fixed effects model

	Estimated parameter		Estimated parameter
ln(RE)	-0.0264	ln(RE)ln(w_L)	-0.0467^{***}
	(0.0743)		(0.0113)
ln(CI)	0.1663^{*}	ln(RE)ln(w_K)	0.0406^{***}
	(0.0904)		(0.0113)
ln(CON)	0.3158^{***}	ln(RE)ln(w_D)	-0.0061
	(0.0543)		(0.0047)
ln(SEC)	-0.8172^{***}	ln(CI)ln(w_L)	0.0144
	(0.0757)		(0.0119)
ln(w_L)	-0.6660^{***}	ln(CI)ln(w_K)	-0.0030
	(0.0968)		(0.0126)
ln(w_K)	0.7735^{***}	ln(CI)ln(w_D)	-0.0113^{***}
	(0.0881)		(0.0042)
ln(w_D)	0.8925^{***}	ln(CON)ln(w_L)	-0.0186^{***}
	(0.0417)		(0.0061)
ln(RE)^2	0.0096	ln(CON) ln(w_K)	0.0026
	(0.0070)		(0.0061)
ln(CI)^2	-0.0529^{***}	ln(CON) ln(w_D)	0.0160^{***}
	(0.0064)		(0.0028)
ln(CON)^2	0.0084^{***}	ln(SEC) ln(w_L)	0.0961^{***}
	(0.0021)		(0.0111
ln(SEC)^2	-0.0082^{***}	ln(SEC) ln(w_K)	-0.0721^{***}
	(0.0035)		(0.0106)
ln(RE) ln(CI)	0.0486^{***}	ln(SEC) ln(w_D)	-0.0224^{***}
	(0.0109)		(0.0041)
ln(RE) ln(CON)	-0.0670^{***}	ln(Asset)	0.7789^{***}
	(0.0051)		(0.0317)
ln(RE) ln(SEC)	0.0553^{***}	ln(Chargeoff)	-0.0111^{***}
	(0.0079)		(0.0033)
ln(CI) ln(CON)	0.0610^{***}	ln(Allow)	0.0234^{**}
	(0.0050)		(0.0010)
ln(CI) ln(CON)	0.0610^{***}	ln(Allow)	0.0234^{**}
	(0.0050)		(0.0010)
ln(CI) ln(CON)	0.0610^{***}	ln(Allow)	0.0234^{**}
	(0.0050)		(0.0010)
ln(CI) ln(SEC)	-0.0210^{***}	t	-0.0098
	(0.0068)		(0.0001)
ln(CON) ln(SEC)	-0.0113^{***}	t^2	$-2.7e-05^{***}$
	(0.0046)		(8.18e-06)
ln(w_L)^2	0.0350^{***}	age	0.0032^{***}
	(0.0039)		(0.0002)
ln(w_K)^2	-0.0083^{*}	σ_u component	
	(0.0051)		
ln(w_D)^2	0.0292^{***}	OCC district	0.3027^{***}
	(0.0020)		(0.0677)
ln(w_L) ln(w_K)	0.0012	Constant	-4.9410^{***}
	(0.0038)		(0.2833)
ln(w_L) ln(w_D)	-0.0363^{***}		
	(0.0030)		
ln(w_K) ln(w_D)	-0.0071^{***}	Economies of Scale	0.9822
	(0.0016)		(0.0210)

Note: We report parameter estimates for the translog cost function specification assuming heteroscedasticity for the technical inefficiency u_{it} and homogeneity of variance for the error term v_{it}.

*** 1%, ** 1% – 5%, * 5% – 10% of p-value.

Economies of scale is calculated based on the sample means of the variables. 500 bootstraps are used to calculate the standard error.

Table 4: Economies of Scope Estimates

Output pair	Scope estimates
RE-CI	0.7265
RE-CON	0.2250
RE-SEC	1.3026
CI-CON	0.9510
CI-SEC	0.2461
CON-SEC	0.8815

Table 5: Elasticities of Substitution and Demand

Allen-Uzawa Elasticities of Substitution (AES)			
	Labor	Capital	Deposit
Labor	-1.5256		
Capital	1.0122	-2.7255	
Deposit	0.7230	1.0700	-1.5118
Morishima Elasticities of Substitution (MES)			
	Labor	Capital	Deposit
Labor		0.9062	0.8029
Capital	1.0333		1.0494
Deposit	0.8189	0.9461	
Price Elasticities of Demand			
	Labor	Capital	Deposit
Labor	-0.5448		
Capital	0.2798	-0.7535	
Deposit	0.2649	0.3921	-0.5540

Table 6: Parameter estimates for the system of equations

	Estimated parameter	Standard error		Estimated parameter	Standard error
θ_{11}	0.3470^{***}	0.0747	θ_{31}	-0.0411^{*}	0.0226
θ_{12}	-0.0554^{***}	0.0794	θ_{32}	0.0698^{*}	0.0403
θ_{13}	-0.3059^{***}	0.0546	θ_{33}	-0.0205	0.0277
θ_{14}	0.0912^{**}	0.0408	θ_{34}	0.0180^{*}	0.0107
θ_{21}	-0.2220^{***}	0.0929	θ_{41}	-0.1781^{***}	0.0603
θ_{22}	0.4977^{***}	0.0761	θ_{42}	-0.1791^{***}	0.0544
θ_{23}	-0.2536^{***}	0.0531	θ_{43}	-0.2041^{***}	0.0498
θ_{24}	0.0152	0.0395	θ_{44}	0.5824^{***}	0.0550
τ	0.4964^{***}	0.0007			
	Sample mean	Standard deviation			
Managerial inefficiency	0.0210	0.0131			
Technical inefficiency	0.1048	0.0563			

Note: Parameter estimates for the system of nonlinear equations are reported using GMM estimation. Managerial inefficiency and technical inefficiency are calculated at sample means.

Figure 2 presents a time series of economies of scale for three tiers, Top 4 (too-big-to-fail banks), Top 5-10, and Top 11-50. All the tiers experience dis-economies of scale, which indicates decreasing cost efficiency over time. The Top 10 have exploited economies of scale even before the 2007-2008 Financial Crisis. Growth in market shares is closely related to decline in economies of scale. The increased sizes and market shares among the Top 4 are not justified by the diminishing economies of scale. All the Top 10 banks have experienced dis-economies since the Financial Crisis.

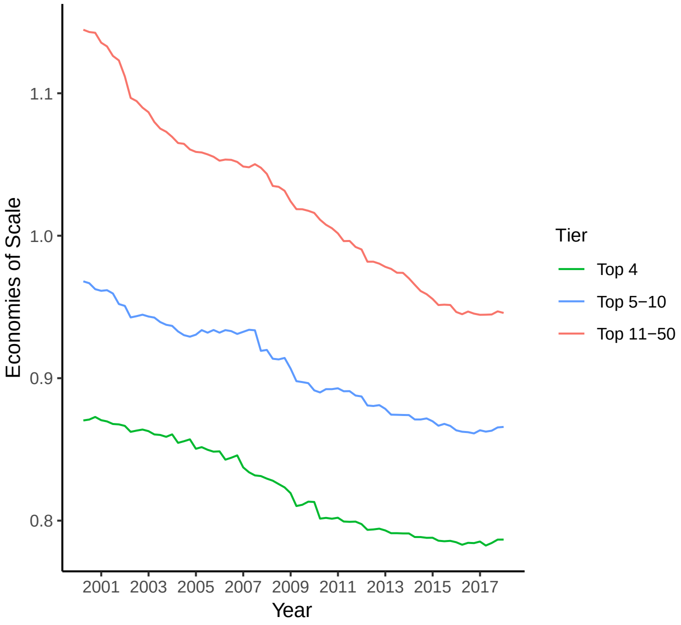

Figure 2: Economies of scale over time

Figure 3 shows technical efficiency by tier and GDP growth over time. The patterns of technical efficiency follows the trend in GDP growth, even though fluctuations may take place at a different time. The major decline in GDP growth which takes place almost two years later than the decrease in technical efficiency during 2007-2008. This can be explained by lagged effect of economic indicators. Similar patterns among the bank tiers show that technical efficiency is closely related to business cycles, indicating a significant impact of environmental factors on technical efficiency.

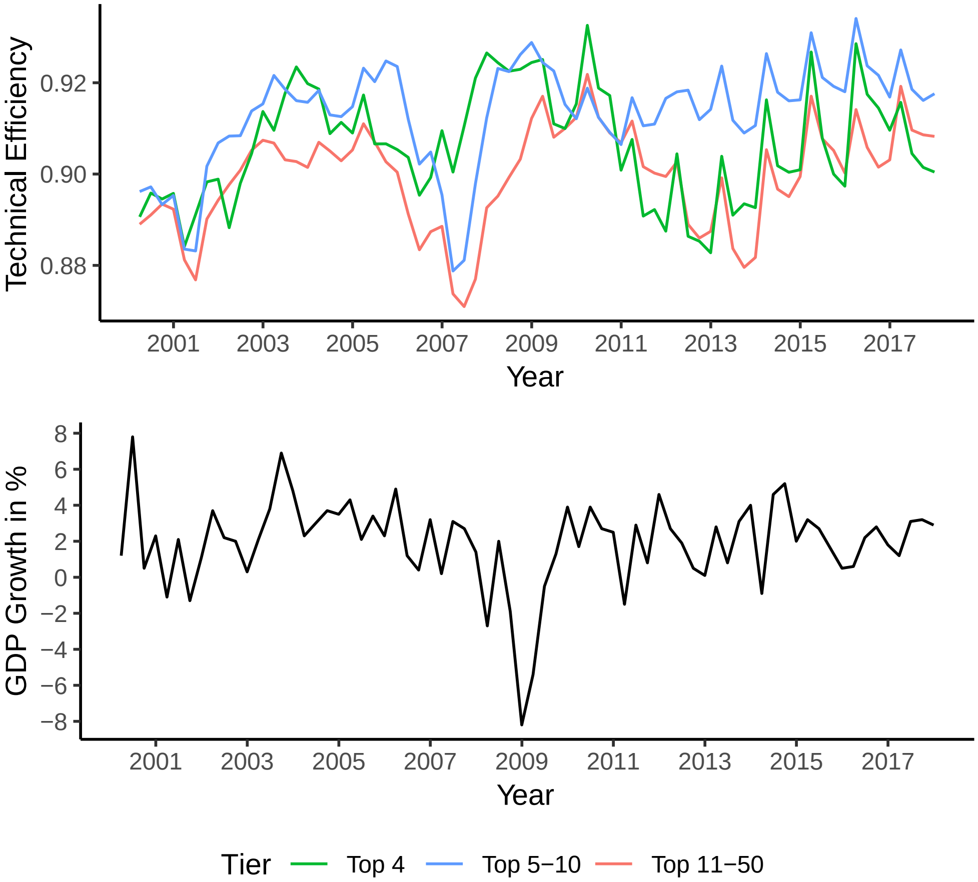

Figure 3: Technical efficiency and GDP growth over time

Figure 4 presents managerial efficiency over time. As in economies of scale, banks of higher ranks have lower managerial efficiency. Banks in the Top 4 are approximately 1.5% and 3.5% less efficient in management than banks in the Top 5-10 and Top 11-50, respectively. These translate into approximately $520 billion and $1.22 trillion of excess cost on average. The average managerial efficiency is almost the same with small fluctuations for lower-rank banks in the sample. Banks in the Top 10 share similar patterns of managerial efficiency, implying similar organizational structure among the banks. The 10 largest banks are more likely to have increased costs of management due to liquidity and capital requirements as well as regulatory policy the government imposes than relatively smaller banks. In fact, the Fed uses $250 billion in total assets or more than $10 billion in foreign exposures on its balance sheet as thresholds, developed more than a decade ago, to define a big bank. A big bank by this definition has to follow, for example, the liquidity coverage ratio, which is used after 2008 and requires banks to have enough cash or liquid assets to cover a month's worth of liabilities in order to reduce the risk of banking collapse. Big banks by this definition are approximately the top 10 banks in our sample. They bear the extra cost from the liquidity and capital rules, and thus are likely to differ in their ability to follow the best-practice banks that do not need to meet the regulations.

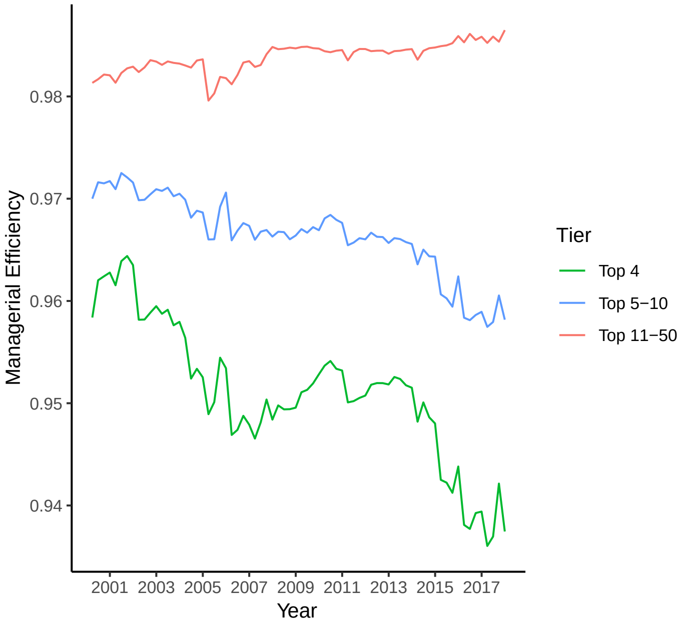

Figure 4: Managerial efficiency over time

Figure 5 shows pooled observations on efficiency. Larger banks tend to have lower managerial efficiency, but there is no definite separation in terms of technical efficiency. Larger banks, however, have smaller variance in technical efficiency.

The decades of various policy changes that resulted in concentrated market power coincide with the increase in pay premium in the banking industry. Figure 6 illustrates the change in the variable components of CEO pay of the top banks. Larger banks tend to pay higher executive compensations. Compared with the other tiers, the Top 4 experience higher

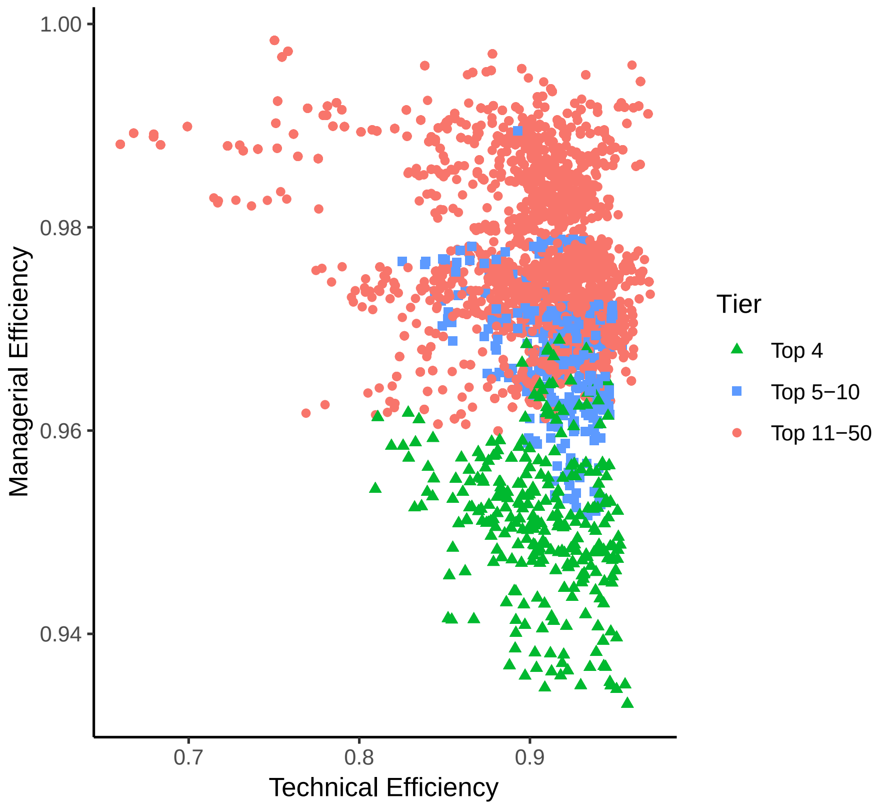

Figure 5: Managerial efficiency vs. technical efficiency

growth in salaries and bonuses after the 2007-2008 Financial Crisis. Value of restricted stock owned for the Top 4 drops to the lowest level before 2011 and soon picks up with other tiers. Value of option awards keeps declining since the early 2000s, and it decreases at a faster rate after 2008. The patterns in the CEO compensations are consistent with findings in Cunat & Guadalupe (2009) which indicate that the level and structure of CEO compensation in the banking industry have changed after the deregulations. We extend the empirical analysis to investigate whether managerial efficiency accounts for the sky-rocketing CEO compensations. Figure 7 presents correlation plots between efficiency and CEO compensations. Managerial efficiency fails to explain the increase in different components of compensations. In fact, higher CEO compensations tend to correlate with lower managerial efficiency. In general, larger banks are more likely to pay higher performance-based incentives but tend to have lower managerial efficiency. The correlations are consistent with findings by Livne, Markarian and Mironov (2013) who show that banks that pay higher incentive plans exhibit more risk and perform worse.

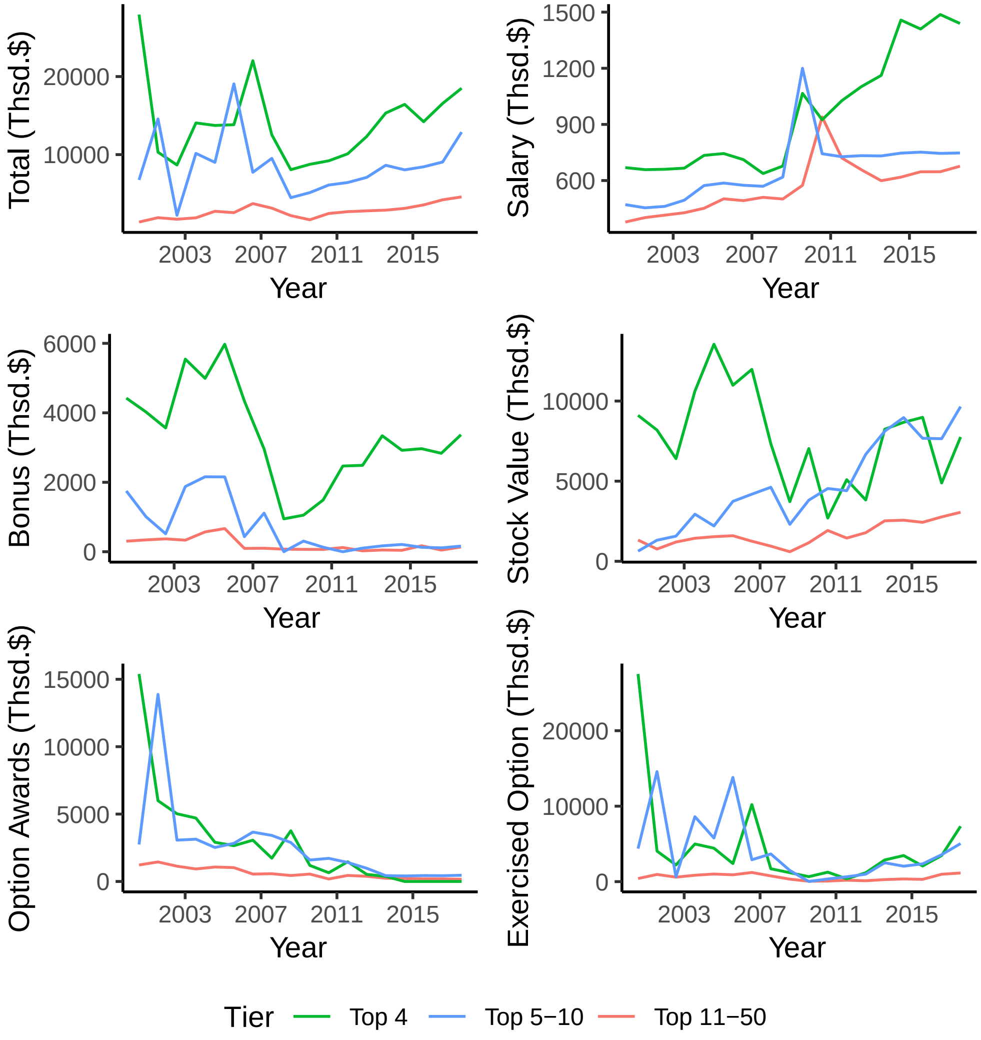

Figure 6: CEO compensation components over time.

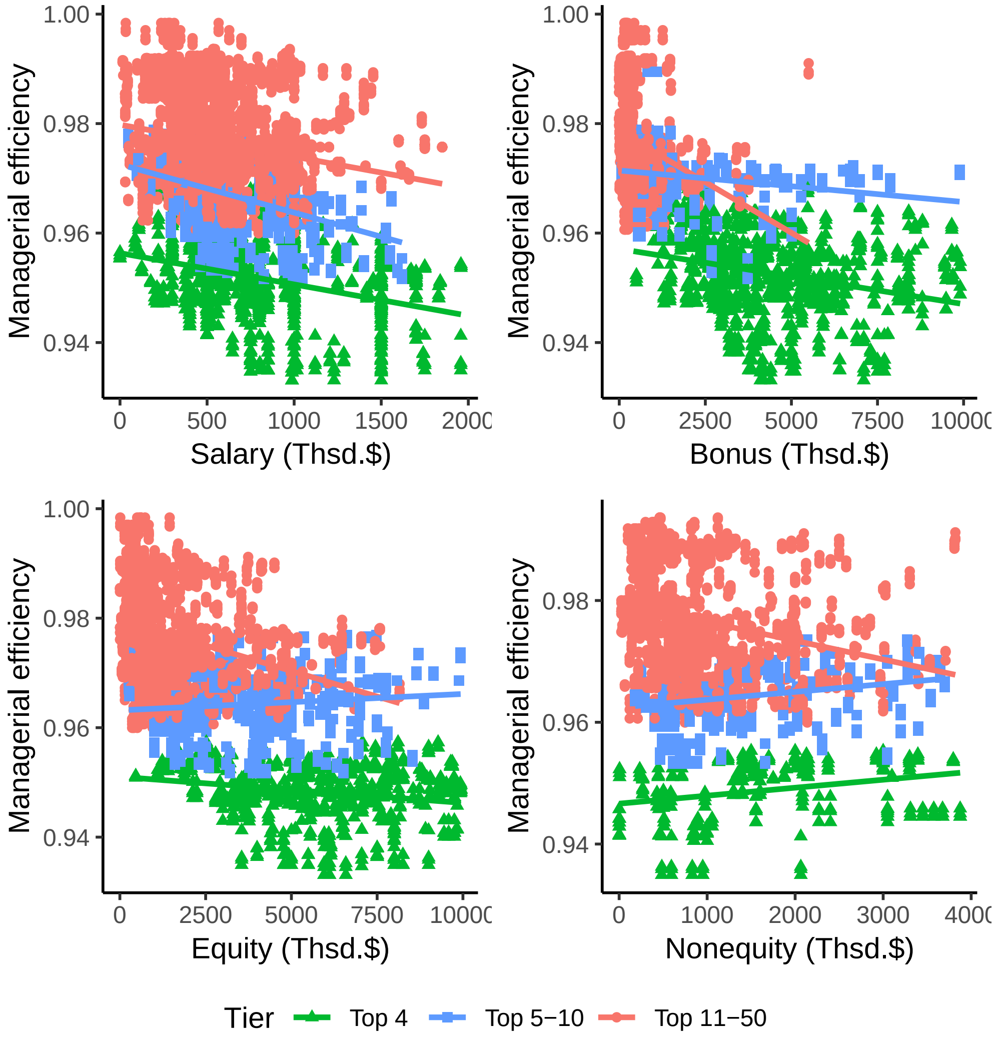

Figure 7: Managerial efficiency vs. CEO compensations

7 Conclusion

We develop a structural model to characterize managerial inefficiency by incorporating managerial decisions, firm effects, and market competition. The essential assumption underlying the model is the existence of the agency problem. Discrepancies exist between the goals of maximizing a manager's utility and achieving the firm/shareholders' goal (i.e. profit maximization). We contribute to the literature on bank performance, market power, and management quality by connecting empirical evidence with an economic model that helps to explain this discrepancy.

We focus on the top 50 U.S. banks, which take up approximately 80% of the U.S. market in terms of total assets from 2000 to 2017 and are systematically important to the U.S. banking industry and the economy. We look at different bank tiers for bank performance in terms of scale economies and efficiency. Scale economies have steadily declined during the time periods for all the bank tiers. This implies that banks have increased in size beyond the minimum efficient scale, and decisions by the managers are at odds with those of the shareholders. Technical efficiency for the sample average follows the GDP growth. It declines during the crisis and starts to pick up with fluctuations since 2015. This suggests that technical efficiency simply picks up excess capacity due to demand shocks and not supply (cost) issues caused by the agency problem. Managerial efficiency declines among the Top 10 but has a slight increase among Top 11-50. The Top 4 are found to have the lowest managerial efficiency. Similar efficiency patterns among the Top 10 imply that the 10 largest banks are likely to bear the extra cost due to regulations on liquidity and capital requirements. Declines in managerial efficiency correspond to diminished economies of scale, but an increase in managerial efficiency does not necessarily correlate with improved scale economies.

We further analyze the correlation between efficiency and variable components of CEO pay. Managerial efficiency fails to explain the increase in the compensations. In general, larger banks are more likely to pay higher performance-based incentives but tend to have lower managerial efficiency. The negative correlation between different pay components and managerial efficiency empirically invalidates the mechanism underlying the incentive plans.

Our modeling efforts in this current research has informed us greatly on a variety of modeling issues that we will leverage in our subsequent modeling efforts. Since the model is based on the assumption that managerial contracts are not optimal, this structural inefficiency also includes the inefficiency from sub-optimal contracts. Therefore, to disentangle the effect of contracts itself would be another direction to develop the research. This modeling effort will endogenize managerial efforts in terms of contracts to model managerial quality.

Compliance with Ethical Standards

Conflict of Interest: Shasha Liu declares that she has no conflict of interest. Robin Sickles declares that he has no conflict of interest.

Ethical approval: This article does not contain any studies with human participants or animals performed by any of the authors.

References

[1] Adams, Robert M., Allen N. Berger, and Robin C. Sickles. 1999. "Semiparametric Approaches to Stochastic Panel Frontier with Applications in the Banking Industry." *Journal of Business & Economic Statistics*, 17, 3: 349-358.

[2] Aigner, Dennis, C.A. Knox Lovell, and Peter Schmidt. 1977. "Formulation and estimation of stochastic frontier production function models." *Journal of Econometrics*, 6,1: 21-37.

[3] Al-Sharkas, Adel A., M. Kabir Hassan, and Shari Lawrence. 2008. "The Impact of Mergers and Acquisitions on the Efficiency of the U.S. Banking Industry: Further Evidence." *Journal of Business Finance & Accounting*, 35, 1-2: 50-70.

[4] Battese, George E., and Tim J. Coelli. 1992. "Frontier production functions, technical efficiency and panel data: With application to paddy farmers in India." *Journal of Productivity Analysis*, 3, 1-2:153-169.

[5] Bauer, Paul W., Allen N. Berger, Gary D. Ferrier, and David B. Humphrey. 1998. "Consistency Conditions for Regulatory Analysis of Financial Institutions: A Comparison of Frontier Efficiency Methods." *Journal of Economics and Business*, 50,2: 85-114.

[6] Baumol, William J., John C. Panzar, and Robert D. Willig. 1982. *Contestable Markets and the Theory of Industry Structure*. New York: Harcourt Brace Jovanovich.

[7] Berger, Allen N., and David B. Humphrey. 1992. "Measurement and efficiency issues in commercial banking." *NBER Chapters: Output measurement in the service sectors*, 245-300.

[8] Charnes, A., W.W. Cooper, and E.Rhodes. 1978. "Measuring the efficiency of decision making units." *European Journal of Operational Research*, 2, 6: 429-444.

[9] Christensen, Laurits R., Dale W. Jorgensen, and Lawrence J. Lau. 1973. "Transcendental Logarithmic Production Frontiers." *Review of Economics and Statistics*, 55, 1: 28-45.

[10] Cornwell, Christopher, Peter Schmidt, and Robin Sickles. 1990. "Production frontiers with cross-sectional and time-series variation in efficiency levels." *Journal of Econometrics*, 46, 1-2: 185-200.

[11] Cunat, Vicente, and Maria Guadalupe. 2009. "Executive compensation and competition in the banking and financial sectors." *Journal of Banking & Finance*, 33, 3: 495-504.

[12] Davies, Richard, and Belinda Tracey. 2014. "Too big to be efficient? The impact of implicit subsidies on estimates of scale economies for banks." *Journal of Money, Credit and Banking*, 46, 1: 219-253.

[13] Debreu, Gerard. 1951. "The coefficient of resource utilization." *Econometrica*, 19(3): 273-292.

[14] Diewert, W. Erwin. 1974. "Applications of Duality Theory". *Frontiers of Quantitive Economics*, 2, ed. M.D. Intrilligator and David Kendrick. Amsterdam: North-Holland Publishing Co.

[15] Drake, Leigh, and Maximilian J.B. Hall. 2003. " Efficiency in Japanese banking: An empirical analysis." *Journal of Banking & Finance*, 27: 891-917.

[16] Farrell, Michael J. 1957. "The Measurement of Productive Efficiency." *Journal of the Royal Statistical Society*, 120, 3: 253-290.

[17] Federal Reserve Statistics. 2017. "Large Commercial Banks." Federal Reserve Statistical Release. https://www.federalreserve.gov/releases/LBR/20171231/default.htm.

[18] Feng, Guohua, and Apostolos Serletis. 2010. "Efficiency, technical change, and returns to scale in large US banks: Panel data evidence from an output distance function satisfying theoretical regularity." *Journal of Banking & Finance*, 34, 1: 127-138.

[19] Ferrier, Garry D., and C.A. Knox Lovell. 1990. "Measuring cost efficiency in banking: Econometric and linear programming evidence." *Journal of Econometrics*, 46, 1-2: 229-245.

[20] Gagnepain, Philippe, and Marc Ivaldi. 2002. "Stochastic Frontiers and Asymmetric Information Models." *Journal of Productivity Analysis*, 18, 2: 145-159.

[21] Greene, William. 2005. "Reconsidering heterogeneity in panel data estimators of the stochastic frontier model." *Journal of Econometrics*, 126: 269-303.

[22] Hall, Alastair R. 2005. *Generalized Method of Moments*, Oxford University Press.

[23] Hughes, Joseph P., Loretta J. Mester, and Choon-Geol Moon. 2001. "Are scale economies in banking elusive or illusive?: Evidence obtained by incorporating capital structure and risk-taking into models of bank production." *Journal of Banking & Finance*, 25, 12 : 2169-2208.

[24] Inanoglu, Hulusi, and Michael Jacobs, Jr. 2009. "Models for Risk Aggregation and Sensitivity Analysis: An Application to Bank Economic Capital." *Journal of Risk Financial Management*, 2,1:118-189.

[25] Inanoglu, Hulusi, Michael Jacobs Jr., Junrong Liu, and Robin Sickles. 2016. "Analyzing Bank Efficiency: Are "Too-Big-to-Fail" Banks Efficient?" *The Handbook of Post Crisis Financial Modeling*, 110-146.

[26] Kutlu, Levent. 2018. "Efficiency estimation in a spatial autoregressive stochastic frontier model." *Economics Letters*, 163: 155-157.

[27] Kutlu, Levent, and Ran Wang. 2018. "Estimation of cost efficiency without cost data." *Journal of Productivity Analysis*, 49, 2-3: 137-151.

[28] Kutlu, Levent, Emmanuel Mamatzakis, and Efthymios G. Tsionas. 2018. "Microfoundations for Performance, Competition and Econometric Implications." Available at https://ssrn.com/abstract=3073430.

[29] Leibenstein, Harvey. 1966. "Allocative Efficiency vs. X-Efficiency." *American Economic Review*, 56: 392-415.

[30] Livne, Gilad, Garen Markarian, and Maxim Mironov. 2013. "Investment horizon, risk, and compensation in the banking industry." *Journal of Banking & Finance*, 37, 9: 3669-3680.

[31] Meeusen, Wim, and Julien van Den Broeck. 1977. "Efficiency estimation from Cobb-Douglas production functions with composed error." *International Economic Review*, 18, 2: 435-444.

[32] Olley, G. Steven, and Ariel Pakes. 1996. "The dynamic of productivity in the telecommunications equipment industry." *Econometrica*, 64, 6: 1263-1297.

[33] Pitt, Mark, and Lung-Fei Lee. 1981. "The measurement and sources of technical inefficiency in the Indonesian weaving industry." *Journal of Development Economics*, 9,1:43-64.

[34] Schmidt, Peter, and Robin Sickles. 1984. "Production Frontiers and Panel Data." *Journal of Business & Economic Statistics*, 2,4: 367-374.

[35] Shephard, Ronald. W. 1953. *Cost and Production Functions. Princeton*, NJ: Princeton University Press.

[36] Simar, Leopold, and Paul W. Wilson. 2013. "Estimation and Inference in Nonparametric Frontier Models: Recent Developments and Perspectives." *Foundations and Trends in Econometrics*, 5, 3-4: 183-337.

[37] Trieb, Thomas P., and Subal C. Kumbhakar. 2018. "Management in production: from unobserved to observed." *Journal of Productivity Analysis*, 49, 2-3: 111-121.

[38] Weill, Laurent. 2004. "Measuring Cost Efficiency in European Banking: A Comparison of Frontier Techniques." *Journal of Productivity Analysis*, 21: 133-152.

[39] Williamson, Oliver. "Managerial Discretion and Business Behavior." *The American Economic Review*, 53, 5: 1032-1057.