

Off-Balance Sheet Activities and Scope Economies in U.S. Banking*

JINGFANG ZHANG¹ EMIR MALIKOV²

¹Auburn University

²University of Nevada, Las Vegas

November 10, 2021

Abstract

Propelled by the recent financial product innovations involving derivatives, securitization and mortgages, commercial banks are becoming more complex, branching out into many “nontraditional” banking operations beyond issuance of loans. This broadening of operational scope in a pursuit of revenue diversification may be beneficial if banks exhibit scope economies. The existing (two-decade-old) empirical evidence lends no support for such product-scope-driven cost economies in banking, but it is greatly outdated and, surprisingly, there has been little (if any) research on this subject despite the drastic transformations that the U.S. banking industry has undergone over the past two decades in the wake of technological advancements and regulatory changes. Commercial banks have significantly shifted towards nontraditional operations, making the portfolio of products offered by present-day banks very different from that two decades ago. In this paper, we provide new and more robust evidence about scope economies in U.S. commercial banking. We improve upon the prior literature not only by analyzing the most recent data and accounting for bank’s nontraditional off-balance sheet operations, but also in multiple methodological ways. To test for scope economies, we estimate a flexible time-varying-coefficient panel-data quantile regression model which accommodates three-way heterogeneity across banks. Our results provide strong evidence in support of significantly positive scope economies across banks of virtually all sizes. Contrary to earlier studies, we find no empirical corroboration for scope diseconomies.

Keywords: bank, cost subadditivity, nontraditional banking, off-balance sheet, product scope, scope economies

JEL Classification: G21, L25, D24

*We are most grateful to the editor, Thorsten Beck, and two anonymous referees for many helpful comments and suggestions. Emir Malikov also would like to acknowledge financial support from the Troesh Center for Entrepreneurship & Innovation at UNLV.

Email: jzz0080@auburn.edu (Zhang) and emir.malikov@unlv.edu (Malikov).

1 Introduction

Just like in other industries, executive managers in banking must choose the optimal scope of operations. Despite the long-lasting implications of this strategic choice for firm performance, the dichotomy between operational “focus” and breadth remains unsettled from the corporate strategy perspective. The common arguments for limited-scope operations à la Skinner (1974a,b) usually feature cost and quality benefits associated with more specialized expertise and tacit knowledge, lessened complexity, diminished technological uncertainty, etc. On the other hand, there may be a strong incentive to diversify revenue streams by broadening the firm’s product mix in order to capitalize on potential scope-driven cost savings and thereby increase firm value (see Panzar and Willig, 1981; Rumelt, 1982; Villalonga, 2004). When it comes to commercial banking, leveraging operational scope and breadth thereof continues to play a vital role in operations management.

The scope of bank operations has also been a subject of intense policy debate, thereby expanding practical importance of understanding the relation between operational scope and bank performance beyond industry managers and stakeholders. Namely, the financial crisis of 2007–2008 and the ensuing Great Recession turned attention of policy-makers and academics alike onto large “too-big-to-fail” (TBTF) commercial banks and the serious systemic risks that they pose. The emergence of behemoth banks due to deregulation as well as technological innovations (including those in information technologies) has given rise to concerns about the costs that such “systemically important financial institutions” impose on the economy and fueled policy debates about whether banks should be subject to size limitations, even including the talks of break-up. These policy discussions have led to the enactment of new financial regulations such as the Dodd–Frank Wall Street Reform and the Consumer Protection Act of 2010 that seek to eliminate the TBTF doctrine by setting restrictions on the scale and scope of bank operations. However, the potential cost savings associated with operating at a large scale with a more diversified scope of revenue-generating activities, which are to be forgone owing to the new regulations, have been by and large neglected in these policy discussions.

Large banks may derive such cost efficiency benefits from their ability to offer financial services at lower average cost due to (i) “scale economies” driven by the increasing returns to scale as well as (ii) their unique position to innovate and expand the scope of offered financial products and thereby economize costs (“scope economies”) via input complementarities and positive spillovers (see Markides and Williamson, 1994; Milgrom and Roberts, 1995) as well as, in the case of commercial banking, risk diversification across different products (e.g., Rossi et al., 2009). In theory, these cost savings are passed onto customers in the form of lower net interest margins. This raises an important policy and research question about significance of the trade-off between lower systemic risk pursued by the newly enacted regulations and the cost savings that banks may be forced to forgo as a result. Both have non-negligible implications for consumer welfare. It is therefore imperative to investigate the prevalence of scale and scope economies in banking in order to not only shed light on potential unintended consequences of the financial reforms already put in place but also to inform future policies and regulations. This information also can help banks in formulating optimal product-scope operational strategies.

While studies of scale economies in commercial banking are many, the attempts to measure *scope economies* are however scant and outdated. The latter is especially lacking given the introduction of many “nontraditional” financial product innovations involving derivatives, securitization and mortgages by the large banks in the past two decades that have allowed them to expand the scope of their revenue-earning operations. The objective of this paper is to fill in this gap.

Early studies of scale economies in banking date as far back as Berger et al. (1987), Mester (1987, 1992) and Hughes and Mester (1993, 1998) to name a few, and with the passage of new financial reforms, this body of research has only been growing. No matter the methods employed, most recent studies find empirical evidence in support of the statistically significant increasing returns to scale in the U.S. banking sector. Some find significant scale economies mostly for large commercial banks (e.g., Wheelock and Wilson, 2012; Hughes and Mester, 2013; Restrepo-Tobón and Kumbhakar, 2015); others find economies of scale for medium and small banks as well (e.g., Malikov et al., 2015; Restrepo-Tobón et al., 2015; Wheelock and Wilson, 2018).

With the sole exception,¹ there however have been virtually no attempt to investigate product scope economies in banking over the past two decades despite the drastic transformations that this sector has undergone during that time. This perhaps can be attributed to the lack of empirical evidence in support of statistically and/or economically significant scope economies among U.S. commercial banks documented in the 1980s and 1990s; e.g., see Berger et al. (1987), Mester (1987), Hughes and Mester (1993), Pulley and Braunstein (1992), Ferrier et al. (1993), Pulley and Humphrey (1993), Jagtiani et al. (1995), Jagtiani and Khanthavit (1996), Wheelock and Wilson (2001). It makes scope economies in the present-day banking sector be a seriously overlooked issue because the technological advancements along with regulatory changes have restructured the U.S. banking industry dramatically, especially since the passage of the Gramm–Leach–Bliley Act in 1999, which largely lifted the restrictions prohibiting the consolidation of commercial banks, investment banks, securities firms and insurance companies. U.S. banks have since experienced a drastic shift from traditional banking activities (viz., issuance of loans) towards the nontraditional activities such as investment banking, venture capital, security brokerage, insurance underwriting and asset securitization (DeYoung and Torna, 2013), and the portfolio of products offered by the modern banks is very different from that two decades ago, underscoring the importance of our study.

While nontraditional banking operations are usually associated with banks’ all other non-interest fee-generating activities related to participating in capital markets, the off-balance sheet banking represents one of the major forms of such nontraditional activities. It chiefly consists of contingent claims/contracts that involve obligations to lend or provide funds should the contingency be realized and, unlike the traditional interest-income-centered transactions, these off-balance sheet activities are not recorded on the bank’s balance sheet (Hassan, 1993; Hassan and Sackley, 1994). For example, an interest-earning loan is considered an asset on the bank’s balance sheet, whereas a promise to make a loan is an off-balance sheet item since it involves only a *potential* funding obligation in the future, albeit, for which

¹To our knowledge, Yuan and Phillips (2008) who explicitly recognize the role of nontraditional banking activities (namely, insurance) is the only attempt at measuring scope economies in the U.S. banking post 2000. Their analysis looks at a single nontraditional operation and stops at 2005, which obviously excludes the most relevant period after the structural-change-inducing financial crisis.

the bank earns a fee. Broadly, off-balance sheet items can be categorized into four groups including guarantees, commitments, market-related activities, and advisory or management functions (e.g., see Perera et al., 2014). Such off-balance sheet banking operations are well-documented to substantially influence banks' financial performance including profitability and risk profiles (e.g., Stiroh, 2004; Laeven and Levine, 2007; Apergis, 2014), and omitting these revenue-earning operations in the analysis of banking technology may lead to erroneous inference and conclusions due to misspecification (see Clark and Siems, 2002; Rime and Stiroh, 2003; Casu and Girardone, 2005; Lozano-Vivas and Pasiouras, 2010). When testing for scope economies, we therefore recognize off-balance sheet operations as another one of the bank's revenue-generating outputs.

In this paper, we contribute to the literature by providing new and more robust evidence about scope economies in U.S. commercial banking. We improve upon the prior literature not only by analyzing the most recent and relevant data (2009–2018) and accounting for bank's nontraditional non-interest-centered operations, but also in multiple methodological ways as follows. In a pursuit of robust estimates of scope economies and statistical inference thereon, we estimate a flexible, yet parsimonious, time-varying-coefficient panel-data quantile regression model which accommodates *(i)* distributional heterogeneity in the cost structure of banks along the size of their costs, *(ii)* temporal variation in cost complementarities and spillovers due to technological change/innovation, and *(iii)* unobserved bank heterogeneity (e.g., latent management quality) that, if unaccounted, confounds the estimates. Our analysis is structural in that we explicitly estimate a model of bank cost structure which facilitates the measurement of counterfactual costs necessary to test for scope economies.

By employing a quantile approach, we are able to capture distributional heterogeneity in the bank cost structure. Unlike the traditional regression models that focus on the conditional mean only, quantile regression provides a complete description of the relationship between the distribution of bank costs and its determinants. Since banks of varying size/scale are highly heterogeneous in their operations (e.g., see Wheelock and Wilson, 2012), it is reasonable to expect that large- and small-scale banks exhibit different scope-driven potential for cost saving (if any) and, therefore, there remains much untapped benefit of examining scope economies in banking via quantile analysis. Thus, contrary to all prior studies of scope economies in banking which provide evidence solely for *average* costs via conventional conditional-mean regressions, we focus our analysis on conditional *quantiles* of the bank cost distribution, with the bank's operating cost being a good proxy for its size/scale. Not only does this approach enable us to accommodate potential heterogeneity in the prevalence of scope economies among banks of different sizes, but it is also more robust to the error distributions including the presence of outliers in the data. Furthermore, it exhibits a useful equivariance property thereby letting us avoid biases in the scope economies computations that numerous earlier studies suffer from (to be discussed later).

To operationalize our analysis, we employ the recently developed quantile estimator (Machado and Santos Silva, 2019) that we extend to allow temporal variation of unknown form in the parameters in order to flexibly capture the impact of technological innovations on bank operations and costs. Our empirical results provide strong evidence in support of statistically significant scope economies across banks virtually of all sizes in the U.S. banking sector. Among banks between the bottom 10th and top 90th

percentiles of the cost distribution, 92% or more exhibit positive economies of scope. The prevalence of significant scope economies in median banks is 99%. Even under the alternative model specifications that produce smaller point estimates, the evidence in support of scope economies in U.S. banking remains strong, with at least 89% of mid-cost banks found to enjoy product-scope-driven cost savings. We also find no empirical corroboration for scope *diseconomies*. Overall, our findings are in stark contrast with earlier studies.

The rest of the paper unfolds as follows. Section 2 discusses the theoretical framework. Section 3 describes our econometric model. Data are discussed in Section 4, followed by Section 5 that reports the empirical results. We then conclude in Section 6.

2 Theory of Multi-Product Costs

In order to test if there is an untapped cost savings potential for commercial banks due to scope economies, we need to formally model their cost structure. Following the convention in the banking literature, we do so using the dual cost approach. Not only is this approach convenient because it facilitates the direct measurement of the bank's costs via the estimated dual cost function necessary for testing for scope economies, but it also does not require the use of input quantities during the estimation (unlike in the primal production approach) which can lead to simultaneity problems since input allocations are the bank's endogenous decision whereas input prices are widely accepted as being exogenously determined owing to competition in the factor market including that for deposits.

A model of bank costs calls for specification of the outputs and inputs of bank production. Given the bank's core functions as a financial intermediary, most studies in the literature adopt Sealey and Lindley's (1977) "intermediation approach" which focuses on the bank's production of intermediation services and the associated costs inclusive of both the interest and operating expenses. In this paradigm, the revenue-generating financial assets such as loans and trading securities are conceptualized as outputs, whereas inputs are typically specified to include labor, physical capital, deposits and other borrowed funds as well as equity capital (for an excellent review, see Hughes and Mester, 2015). Given the recent industry trends and the growing importance of nontraditional income-earning activities that banks engage in, we also include an output measure of non-interest off-balance sheet income. Together with loans and securities, this makes a total of $M = 3$ outputs.

Concretely, we formalize the bank's cost structure via the following multi-product dual variable cost function:

$$\mathcal{C}_t(\mathbf{Y}, \mathbf{W}, \mathbf{K}) = \min_{\mathbf{X} \geq \mathbf{0}} \{ \mathbf{X}' \mathbf{W} \mid (\mathbf{X}, \mathbf{K}) \text{ can produce } \mathbf{Y} \text{ at time } t \}, \quad (2.1)$$

where the arguments of cost function $\mathcal{C}_t(\cdot)$ are the output quantities $\mathbf{Y} \in \mathbb{R}_+^M$, variable input prices $\mathbf{W} \in \mathbb{R}_+^J$ and fixed input quantities $\mathbf{K} \in \mathbb{R}_+^P$; and $\mathbf{X} \in \mathbb{R}_+^J$ is the vector of variable input quantities. Importantly, the cost function in (2.1) is time-varying thereby accommodating the evolution of the bank cost structure over time in the face of technological advancements and regulatory changes.

The multi-product firm's cost structure is said to exhibit scope economies if its average cost is de-

creasing in the number of outputs/operations (Panzar and Willig, 1981). Commercial banks may achieve such cost savings by spreading fixed costs (e.g., branch costs and data processing costs) over the more diversified output mix (fixed asset amortization) which now, more often than not, includes nontraditional off-balance sheet operations. Scope economies may also arise from positive spillovers via the (re)use of “public inputs” such as client credit information and customer relations as well as intangible assets including tacit knowledge and know-hows. Complementarities across different products can play a big role too. For example, some off-balance sheet operations such as loan commitments (which generate income for banks via fees) essentially represent a technological expansion of traditional lending at a little cost added. At the same time, they can help banks expand the scope of their customer relationship with all the cost-saving informational gains that come with it (Berger and Udell, 1995; Das and Nanda, 1999; Degryse and Van Cayseele, 2000). Banks can also reuse the information gathered when issuing loans to reduce the searching or monitoring requirements of the off-balance sheet activities.

To test for the potential for scope-driven cost savings, we use an expansion-path measure of subadditivity of the bank’s cost function à la Berger et al. (1987), with the rationale being that subadditivity sheds light on scope economies, the presence of which is a necessary condition for the former (see Baumol et al., 1982; Evans and Heckman, 1984). Specifically, the subadditivity measure relies on comparison of the costs of smaller *multi*-output banks of *differential* degrees of specialization with the cost of a larger, more diversified bank.² Intuitively, this approach zeroes in on scope economies from a perspective of relative—as opposed to absolute—notation of revenue diversification. Then, for some distribution weights $0 \leq \omega_m^\kappa \leq 1$ such that $\sum_\kappa \omega_m^\kappa = 1$ for all $m = 1, 2, 3$ and $\kappa \in \{A, B, C\}$, the bank is said to enjoy scope economies at time t if

$$\sum_{\kappa \in \{A, B, C\}} \mathcal{C}_t(\omega_1^\kappa Y_1, \omega_2^\kappa Y_2, \omega_3^\kappa Y_3) - \mathcal{C}_t(Y_1, Y_2, Y_3) > 0, \quad (2.2)$$

where we have suppressed all arguments of the cost function besides outputs.

While the above methodology deviates from the conventional definition of scope economies (Baumol et al., 1982) which relies on the comparison of the cost of producing outputs individually with the cost of their joint production, whereby the bank is said to enjoy scope economies if $\mathcal{C}_t(Y_1, 0, 0) + \mathcal{C}_t(0, Y_2, 0) + \mathcal{C}_t(0, 0, Y_3) - \mathcal{C}_t(Y_1, Y_2, Y_3) > 0$, it is both more realistic and robust. This is so because it does not require computation of the counterfactual cost of producing each output separately by a fully specialized *single*-output bank, which naturally suffers from “excessive extrapolation” (Evans and Heckman, 1984; Hughes and Mester, 1993) since the counterfactuals require extrapolation of the estimated multi-output cost function to its boundaries corresponding to the *non-existent* single-output specializations. Also, the conventional measure of scope economies is just a special case of (2.2) with a pair of weights taking zero values for each counterfactual bank.

To further avoid excessive extrapolation, we restrict the choice of $\{\omega_m\}$ to the “admissible region” defined by the two data-driven constraints, following Evans and Heckman (1984). First, each counterfactual bank is ensured to not produce less of each output than banks do in the sample. That is, we require that $\omega_m^\kappa Y_m \geq \min\{Y_m\}$ for all $m = 1, 2, 3$ and $\kappa \in \{A, B, C\}$. The second constraint ensures that

²While preserving the equality of total output quantities on both sides, of course.

each counterfactual bank does not specialize in either one of the outputs to a greater extent than banks do in the sample. In other words, ratios of output quantities for each counterfactual bank must fall in the range of such ratios observed in the data, i.e., for any pair Y_m and $Y_{m'}$:

$$\min \left\{ \frac{Y_m}{Y_{m'}} \right\} \leq \frac{\bar{\omega}_m^\kappa Y_m^* + \min\{Y_m\}}{\bar{\omega}_{m'}^\kappa Y_{m'}^* + \min\{Y_{m'}\}} \leq \max \left\{ \frac{Y_m}{Y_{m'}} \right\}, \quad (2.3)$$

where $Y_m^* = Y_m - 3 \times \min\{Y_m\}$ for all $m = 1, 2, 3$. Thus, we examine the *within-sample* scope economies.

The quantitative measure of cost subadditivity $\mathcal{S}_t$ (in proportions) is obtained by dividing the expression in (2.2) by $\mathcal{C}_t(Y_1, Y_2, Y_3)$:

$$\mathcal{S}_t = \frac{\sum_{\kappa \in \{A, B, C\}} \mathcal{C}_t \left(\bar{\omega}_1^\kappa Y_1^* + \min\{Y_1\}, \bar{\omega}_2^\kappa Y_2^* + \min\{Y_2\}, \bar{\omega}_3^\kappa Y_3^* + \min\{Y_3\} \right) - \mathcal{C}_t(Y_1, Y_2, Y_3)}{\mathcal{C}_t(Y_1, Y_2, Y_3)}, \quad (2.4)$$

where the counterfactual costs under the summation operator have been redefined in order to operationalize the first of the two constraints characterizing the admissible region. Positive (negative) values of $\mathcal{S}_t$ provide evidence of scope economies (*diseconomies*); while a zero value suggests scope invariance of the bank's cost structure.

Clearly however, the value of $\mathcal{S}_t$ depends on the choice of distribution weights $\{\bar{\omega}_m^\kappa\}$. To test for scope economies, we adopt a conservative approach to measuring cost subadditivity, whereby $\{\bar{\omega}_m^\kappa\}$ are chosen such that the corresponding $\mathcal{S}_t$ is the smallest. With this, “the” measure of cost subadditivity (for each bank-year) is

$$\mathcal{S}_t^* = \min_{\{\bar{\omega}_m^\kappa\}} \mathcal{S}_t(\bar{\omega}_m^\kappa; m = 1, 2, 3; \kappa \in \{A, B, C\}). \quad (2.5)$$

The rationale is as follows. If the *smallest* subadditivity measure is still positive, then one can quite safely infer that scope economies are locally significant over the bank's feasible output space in a given year. Thus, the main hypothesis of interest is as follows.

HYPOTHESIS.—*Consistent with scope economies at time t , the cost subadditivity measure $\mathcal{S}_t^* > 0$.*

3 Empirical Model

We estimate the bank's dual variable cost function $\mathcal{C}_t(\cdot)$ at different conditional quantiles of costs. Let C_{it} be the variable cost of a bank $i = 1, \dots, n$ in year $t = 1, \dots, T$ and $\mathbf{V}_{it} = (\mathbf{Y}_{it}', \mathbf{W}_{it}', \mathbf{K}_{it}')'$ be the vector of (strictly exogenous) cost-function regressors. We use lower case of C_{it} and $\mathbf{V}_{it}$ in the following to denote the log transformations of the variables: e.g., $\mathbf{v}_{it} = \ln \mathbf{V}_{it}$. Letting the bank's variable cost structure be of the translog³ form and described by a location-scale model à la Koenker and Bassett (1982) extended to accommodate bank fixed effects and time-varying coefficients, we have

$$c_{it} = [\beta_0 + \beta_0^* L(t)] + [\boldsymbol{\beta}_1 + \boldsymbol{\beta}_1^* L(t)]' \mathbf{v}_{it} + \frac{1}{2} [\boldsymbol{\beta}_2 + \boldsymbol{\beta}_2^* L(t)]' \text{vec}(\mathbf{v}_{it} \mathbf{v}_{it}') + \lambda_i + u_{it}, \quad (3.1)$$

with

$$u_{it} = \left([\gamma_0 + \gamma_0^* S(t)] + [\boldsymbol{\gamma}_1 + \boldsymbol{\gamma}_1^* S(t)]' \mathbf{v}_{it} + \frac{1}{2} [\boldsymbol{\gamma}_2 + \boldsymbol{\gamma}_2^* S(t)]' \text{vec}(\mathbf{v}_{it} \mathbf{v}_{it}') + \sigma_i \right) \varepsilon_{it}, \quad (3.2)$$

³Quadratic log-polynomial.

where $(\beta_0, \beta'_1, \beta'_2, \beta_0^*, \beta_1^{*'}, \beta_2^{*'})'$ are unknown location-function coefficients; $(\gamma_0, \gamma'_1, \gamma'_2, \gamma_0^*, \gamma_1^{*'}, \gamma_2^{*'})'$ are unknown scale-function coefficients; and λ_i and σ_i are the unobserved bank-specific location and scale fixed effects, respectively.

To allow for technological change in the bank cost structure, we borrow from Baltagi and Griffin (1988) and introduce two scalar time indices $L(t)$ and $S(t)$. Both time indices are unobservable and can be thought of as the unknown functions of time. Such time indices are advantageous over simple trends (including quadratic) in modeling temporal changes because they provide richer variation in the measurement of technological change and much closer approximation to observed temporal changes than do the simple time trends. Note that index $L(t)$ enters the location function non-neutrally, shifting not only the intercept $\beta_0 + \beta_0^* L(t)$ but also the linear $\beta_1 + \beta_1^* L(t)$ and quadratic slopes $\beta_2 + \beta_2^* L(t)$, thereby allowing for flexible locational shifts in the costs over time. Analogous scale changes over time are allowed by means of $S(t)$. In all, by means of the time indices in both the location and scale functions, we are able to accommodate temporal changes in the *entire* conditional cost distribution.

Essentially, our model in (3.1)–(3.2) is a generalization of the popular translog cost-function specification, where all parameters now vary with time, the covariates affect not only the location (centrality) but also the scale (variability) of the conditional cost distribution; and the bank fixed effects are both location- and scale-shifting. The two equations together facilitate a quantile analysis of the bank's cost structure. Along the lines of Machado and Santos Silva (2019) upon whom we build our estimation procedure, we assume that (i) ε_{it} is *i.i.d.* across i and t with some cdf F_ε ; (ii) $\varepsilon_{it} \perp \mathbf{v}_{it}$ with the normalizations that $\mathbb{E}[\varepsilon_{it}] = 0$ and $\mathbb{E}[|\varepsilon_{it}|] = 1$; and (iii) $\Pr \left[[\gamma_0 + \gamma_0^* S(t)] + [\gamma_1 + \gamma_1^* S(t)]' \mathbf{v}_{it} + \frac{1}{2} [\gamma_2 + \gamma_2^* S(t)]' \times \text{vec}(\mathbf{v}_{it} \mathbf{v}_{it}') + \sigma_i > 0 \right] = 1$. Then, for any given quantile index $\tau \in (0, 1)$, the τ th conditional quantile function of the log-cost c_{it} implied by (3.1)–(3.2) is

$$\begin{aligned} \mathcal{Q}_c[\tau | \mathbf{v}_{it}] = & \underbrace{\left[\beta_0 + \gamma_0 q_\tau + \beta_0^* L(t) + \gamma_0^* S(t) q_\tau \right]}_{t\text{-varying quantile intercept}} + \underbrace{\left[\beta_1 + \gamma_1 q_\tau + \beta_1^* L(t) + \gamma_1^* S(t) q_\tau \right]'}_{t\text{-varying linear quantile slopes}} \mathbf{v}_{it} + \\ & \underbrace{\frac{1}{2} \left[\beta_2 + \gamma_2 q_\tau + \beta_2^* L(t) + \gamma_2^* S(t) q_\tau \right]'}_{t\text{-varying quadratic quantile slopes}} \text{vec}(\mathbf{v}_{it} \mathbf{v}_{it}') + \underbrace{\left[\lambda_i + \sigma_i q_\tau \right]}_{\text{individual quantile fixed effect}}, \end{aligned} \quad (3.3)$$

where $q_\tau = F_\varepsilon^{-1}(\tau)$ is the (unknown) τ th quantile of ε_{it} .

The translog cost model in (3.3) is quantile-specific because all bracketed “composite” coefficients vary not only with time but also with the cost quantile τ . Furthermore, the technological change in the cost frontier is also quantile-specific thereby allowing for heterogeneous temporal shifts across the entire cost distribution as opposed to a shift in the mean only. The unobserved bank fixed effect inside the last brackets is also quantile-specific. Thus, quantile model (3.3) can be rewritten compactly as

$$\mathcal{Q}_c[\tau | \mathbf{v}_{it}] \equiv \alpha_0(\tau, t) + \boldsymbol{\alpha}_1(\tau, t)' \mathbf{v}_{it} + \frac{1}{2} \boldsymbol{\alpha}_2(\tau, t)' \text{vec}(\mathbf{v}_{it} \mathbf{v}_{it}') + \mu_{i,\tau}, \quad (3.4)$$

with the “alpha” coefficients corresponding to the bracketed expressions in (3.3) and $\mu_i \equiv \lambda_i + \sigma_i q_\tau$.

We opt to begin with the location-scale model to derive the conditional quantile function of interest in (3.3) as opposed to postulating a quantile regression à la (3.4) *prima facie* because we seek to estimate these quantiles *indirectly*. This is motivated by the presence of unobserved fixed effects in the quantile

model. Namely, since there is no known general transformation that can purge unit fixed effects from the quantile model (owing to nonlinearity of the quantile operator), in such a case the routine check-function-based estimators proceed to *directly* estimate a vector of individual effects by means of including a full set of unit dummies. However, as noted by Koenker (2004), the introduction of a large number of unit fixed effects significantly inflates the variability of estimates of the main parameters of interest, i.e., the slope coefficients. Furthermore, the optimization of an L_1 -norm corresponding to the check-function-based estimators, when there is a large number of binary variables and the associated parameters to be estimated, is well-known to be computationally cumbersome and oftentimes intractable in practice.⁴ The traditional solution to this assumes that unit fixed effects are only location-shifting and regularizes these individual effects by shrinking them to a common value (see Koenker, 2004; Lamarche, 2010), but these estimators have gained little popularity in applied work largely because of their complexity. While there is an alternative fixed-effect quantile estimator proposed by Canay (2011) that requires no regularization and is notably simpler to implement, it continues to assume that the unit fixed effects have a pure location shift effect. Using the notation of (3.4), this is tantamount to assuming that $\mu_{i,\tau} = \mu_i$ for all τ . Furthermore, none of these check-function-based estimators guarantee that the estimates of regression quantiles do not cross, which is a pervasive but oft-ignored problem in applied work. We therefore adopt the approach recently proposed by Machado and Santos Silva (2019) that allows an easy-to-implement *indirect* estimation of the quantile parameters via moments, where all parameters are estimated based on the moments implied by the location-scale model in (3.1)–(3.2). Besides its relative computational simplicity, this approach is advantageous for its ability to control for unobserved unit heterogeneity that is both location- and scale-shifting: the individual effects are allowed to affect the entire distribution rather than just shifting its location (therefore, $\{\mu_{i,\tau}\}$ are also quantile-specific). Lastly but not least importantly, this moment-based approach can be easily applied to nonlinear-in-parameters models (like ours is) and produces non-crossing quantile regressions.

To operationalize the estimator, we model unobservable $L(t)$ and $S(t)$ via discretization. For each $\kappa = 1, \dots, T$, define the dummy variable $D_{\kappa,t}$ that is equal to 1 in the κ th time period and 0 otherwise. Then, we discretize time indices as $L(t) = \sum_{\kappa=2}^T \eta_{\kappa} D_{\kappa,t}$ and $S(t) = \sum_{\kappa=2}^T \theta_{\kappa} D_{\kappa,t}$, where $L(1) = \eta_1 = 0$ and $S(1) = \theta_1 = 0$ are normalized for identification. Parameter identification also requires that both β_0^* and γ_0^* be normalized; we set $\beta_0^* = \gamma_0^* = 1$. Under these identifying normalizations, β_0 , β_1 , γ_0 and γ_1 are naturally interpretable as “reference” coefficients in time period $t = 1$. Then, a feasible analogue of the τ th conditional cost quantile in (3.3) is given by

$$Q_c[\tau | \mathbf{v}_{it}] = \left[\beta_0 + \gamma_0 q_{\tau} + \sum_{\kappa} (\eta_{\kappa} + \theta_{\kappa} q_{\tau}) D_{\kappa,t} \right] + \left[\beta_1 + \gamma_1 q_{\tau} + \sum_{\kappa} (\beta_1^* \eta_{\kappa} + \gamma_1^* \theta_{\kappa} q_{\tau}) D_{\kappa,t} \right]' \mathbf{v}_{it} + \frac{1}{2} \left[\beta_2 + \gamma_2 q_{\tau} + \sum_{\kappa} (\beta_2^* \eta_{\kappa} + \gamma_2^* \theta_{\kappa} q_{\tau}) D_{\kappa,t} \right]' \text{vec}(\mathbf{v}_{it} \mathbf{v}_{it}') + [\lambda_i + \sigma_i q_{\tau}]. \quad (3.5)$$

Two remarks are in order. First, the discretized parameterization of the unknown $L(t)$ and $S(t)$ is akin to a nonparametric local-constant estimation of these unknown functions of time with the bandwidth parameter being set to 0. Second, though it might appear at first that, when $L(t)$ and $S(t)$ are modeled us-

⁴For instance, in our empirical application $n > 7,500$.

ing a series of time dummies, we obtain the time-varying slope coefficients on $\mathbf{v}_{it}$ by merely interacting the latter with time dummies and adding them as additional regressors, this is *not* the case here because time dummies are restricted to have the same parameters $\{\eta_k\}$ and $\{\theta_k\}$ both when entering additively as well as when interacting with $\mathbf{v}_{it}$. Thus, the location and scale functions are not “fully saturated” specification but, in fact, are more parsimonious *nonlinear* (in parameters) functions with much fewer unknown parameters. In avoiding a fully saturated specification that is equivalent to sample-splitting into cross-sections, we accommodate time-invariant bank fixed effects.

3.1 Estimation Procedure

Although the estimation of (3.3) [or (3.5)] can be done in one step via nonlinear method of moments, we adopt a multi-step procedure that is significantly easier to implement. This is possible because the moments implied by model (3.1)–(3.2) and its assumptions are sequential in nature. In other words, we can first estimate parameters of the location function and then those of the scale function in two separate steps. After that, based on the estimates of these parameters, the third step is taken to estimate unknown quantiles and, ultimately, recover time-varying quantile coefficients in (3.3). In what follows, we briefly describe this procedure, with more details available in Appendix A.

Step 1. We first estimate parameters of the location function. For ease of notation, let $\mathbf{D}_t = [D_{2,t}, \dots, D_{T,t}]'$ and $\boldsymbol{\eta} = [\eta_2, \dots, \eta_T]'$. Under the assumption (ii), from (3.1) it follows that the conditional mean function of the log-cost c_{it} is

$$\mathbb{E}[c_{it} | \mathbf{v}_{it}, \mathbf{D}_t] = \beta_0 + \boldsymbol{\eta}' \mathbf{D}_t + [\boldsymbol{\beta}_1 + \boldsymbol{\eta}' \mathbf{D}_t \cdot \boldsymbol{\beta}_1^*]' \mathbf{v}_{it} + \frac{1}{2} [\boldsymbol{\beta}_2 + \boldsymbol{\eta}' \mathbf{D}_t \cdot \boldsymbol{\beta}_2^*]' \text{vec}(\mathbf{v}_{it} \mathbf{v}_{it}') + \lambda_i, \quad (3.6)$$

which can be consistently estimated in the within-transformed form via nonlinear least squares after purging additive location fixed effects. Having obtained the nonlinear fixed-effects estimates of the slope coefficients $(\hat{\boldsymbol{\eta}}', \hat{\boldsymbol{\beta}}_1', \hat{\boldsymbol{\beta}}_1^{*'}, \hat{\boldsymbol{\beta}}_2', \hat{\boldsymbol{\beta}}_2^{*'})'$, we can then recover the location-shifting intercept β_0 and fixed effects $\{\lambda_i\}$ under the usual $\sum_{i=1}^n \lambda_i = 0$ normalization:

$$\hat{\beta}_0 = \frac{1}{nT} \sum_i \sum_t \left(c_{it} - \hat{\boldsymbol{\eta}}' \mathbf{D}_t - [\hat{\boldsymbol{\beta}}_1 + \hat{\boldsymbol{\eta}}' \mathbf{D}_t \cdot \hat{\boldsymbol{\beta}}_1^*]' \mathbf{v}_{it} - \frac{1}{2} [\hat{\boldsymbol{\beta}}_2 + \hat{\boldsymbol{\eta}}' \mathbf{D}_t \cdot \hat{\boldsymbol{\beta}}_2^*]' \text{vec}(\mathbf{v}_{it} \mathbf{v}_{it}') \right), \quad (3.7)$$

$$\hat{\lambda}_i = \frac{1}{T} \sum_t \left(c_{it} - \hat{\beta}_0 - \hat{\boldsymbol{\eta}}' \mathbf{D}_t - [\hat{\boldsymbol{\beta}}_1 + \hat{\boldsymbol{\eta}}' \mathbf{D}_t \cdot \hat{\boldsymbol{\beta}}_1^*]' \mathbf{v}_{it} - \frac{1}{2} [\hat{\boldsymbol{\beta}}_2 + \hat{\boldsymbol{\eta}}' \mathbf{D}_t \cdot \hat{\boldsymbol{\beta}}_2^*]' \text{vec}(\mathbf{v}_{it} \mathbf{v}_{it}') \right) \forall i. \quad (3.8)$$

Hence, the residual is $\hat{u}_{it} = c_{it} - \hat{\beta}_0 - \hat{\boldsymbol{\eta}}' \mathbf{D}_t - [\hat{\boldsymbol{\beta}}_1 + \hat{\boldsymbol{\eta}}' \mathbf{D}_t \cdot \hat{\boldsymbol{\beta}}_1^*]' \mathbf{v}_{it} - \frac{1}{2} [\hat{\boldsymbol{\beta}}_2 + \hat{\boldsymbol{\eta}}' \mathbf{D}_t \cdot \hat{\boldsymbol{\beta}}_2^*]' \text{vec}(\mathbf{v}_{it} \mathbf{v}_{it}') - \hat{\lambda}_i$.

Step 2. We then estimate parameters of the scale function. Based on the assumptions (ii)–(iii), we have an auxiliary conditional mean regression:

$$\mathbb{E}[|u_{it}| | \mathbf{v}_{it}, \mathbf{D}_t] = \gamma_0 + \boldsymbol{\theta}' \mathbf{D}_t + [\boldsymbol{\gamma}_1 + \boldsymbol{\theta}' \mathbf{D}_t \cdot \boldsymbol{\gamma}_1^*]' \mathbf{v}_{it} + \frac{1}{2} [\boldsymbol{\gamma}_2 + \boldsymbol{\theta}' \mathbf{D}_t \cdot \boldsymbol{\gamma}_2^*]' \text{vec}(\mathbf{v}_{it} \mathbf{v}_{it}') + \sigma_i, \quad (3.9)$$

where $\boldsymbol{\theta} = [\theta_2, \dots, \theta_T]'$ and which, just like in the first step, we can estimate via nonlinear least squares after within-transforming scale fixed effects out. This yields the estimates of the scale-function slope coefficients $(\hat{\boldsymbol{\theta}}', \hat{\boldsymbol{\gamma}}_1', \hat{\boldsymbol{\gamma}}_1^{*'}, \hat{\boldsymbol{\gamma}}_2', \hat{\boldsymbol{\gamma}}_2^{*'})'$. To recover the scale-shifting intercept γ_0 and fixed effects $\{\sigma_i\}$, use

$$\sum_{i=1}^n \sigma_i = 0:$$

$$\hat{\gamma}_0 = \frac{1}{nT} \sum_i \sum_t \left(|\hat{u}_{it}| - \hat{\theta}' \mathbf{D}_t - [\hat{\gamma}_1 + \hat{\theta}' \mathbf{D}_t \cdot \hat{\gamma}_1^*]' \mathbf{v}_{it} - \frac{1}{2} [\hat{\gamma}_2 + \hat{\theta}' \mathbf{D}_t \cdot \hat{\gamma}_2^*]' \text{vec}(\mathbf{v}_{it} \mathbf{v}_{it}') \right), \quad (3.10)$$

$$\hat{\sigma}_i = \frac{1}{T} \sum_t \left(|\hat{u}_{it}| - \hat{\gamma}_0 - \hat{\theta}' \mathbf{D}_t - [\hat{\gamma}_1 + \hat{\theta}' \mathbf{D}_t \cdot \hat{\gamma}_1^*]' \mathbf{v}_{it} - \frac{1}{2} [\hat{\gamma}_2 + \hat{\theta}' \mathbf{D}_t \cdot \hat{\gamma}_2^*]' \text{vec}(\mathbf{v}_{it} \mathbf{v}_{it}') \right) \forall i. \quad (3.11)$$

Step 3. For any given quantile index $0 < \tau < 1$ of interest, we next estimate the unconditional quantile of ε_{it} . From (3.2), we have the conditional quantile function of u_{it} :

$$\mathcal{Q}_u[\tau | \mathbf{v}_{it}, \mathbf{D}_t] = \left(\gamma_0 + \theta' \mathbf{D}_t + [\gamma_1 + \theta' \mathbf{D}_t \cdot \gamma_1^*]' \mathbf{v}_{it} + \frac{1}{2} [\gamma_2 + \theta' \mathbf{D}_t \cdot \gamma_2^*]' \text{vec}(\mathbf{v}_{it} \mathbf{v}_{it}') + \sigma_i \right) q_\tau, \quad (3.12)$$

and therefore we can estimate q_τ via the standard quantile regression of $\hat{u}_{it}$ from Step 1 on $(\hat{\gamma}_0 + \hat{\theta}' \mathbf{D}_t + [\hat{\gamma}_1 + \hat{\theta}' \mathbf{D}_t \cdot \hat{\gamma}_1^*]' \mathbf{v}_{it} + \frac{1}{2} [\hat{\gamma}_2 + \hat{\theta}' \mathbf{D}_t \cdot \hat{\gamma}_2^*]' \text{vec}(\mathbf{v}_{it} \mathbf{v}_{it}') + \hat{\sigma}_i)$ from Step 2, with no intercept. With all unknown parameters now estimated, we can construct the estimator of the feasible analogue of the τ th conditional quantile of the log-cost in (3.5).

For statistical inference, we use bootstrap. To correct for finite-sample biases, we employ Efron's (1982) bias-corrected bootstrap percentile confidence intervals. Bootstrap also significantly simplifies testing because, owing to a multi-step nature of our estimator, computation of the asymptotic variance of the parameter estimators is not trivial. Due to the panel structure of data, we use wild residual *block* bootstrap, thereby taking into account the potential dependence in residuals within each bank over time. Details are provided in Appendix B.

4 Data

The bank-level data come from the Reports of Condition and Income (the so-called Call Reports) and the Uniform Bank Performance Reports (UBPRs). We obtain annual year-end data for all FDIC-insured commercial banks between 2009 and 2018. As already discussed at length, we focus on the post-financial-crisis period.

Consistent with the widely accepted Sealey and Lindley's (1977) "intermediation approach" to formalizing production in banking, we define the bank's cost-function arguments as follows. The two traditional interest-income-centered outputs are Y_1 — total loans, which include real estate loans, agricultural loans, commercial and industrial loans, individual consumer loans and other loans, and Y_2 — total securities, which is the sum of securities held-to-maturity and securities held-for-sale. These output categories are conventional and the same as those considered by, e.g., Koetter et al. (2012) and Wheelock and Wilson (2020). The third output included in our analysis (Y_3) measures nontraditional off-balance sheet operations. We use a sum of credit-equivalent measures of the bank's various off-balance sheet operations as a proxy for its involvement in nontraditional activities. Namely, we convert off-balance sheet items into their *credit equivalents* which we determine using credit conversion factors that account for the varying credit risk of different nontraditional banking operations.⁵ This facilitates comparability

⁵For example, financial standby letter of credit and repo-style transactions have a credit conversion factor of 1, whereas performance standby letters of credit have a factor of 0.5. The conversion factors come from the Call Reports.

of (traditional) on- and (nontraditional) off-balance sheet activities in the analysis of banking production, which makes it a popular practice in the literature (e.g., Jagtiani and Khanthavit, 1996; Hughes and Mester, 1998; Stiroh, 2000; Clark and Siems, 2002; Berger and Mester, 2003; Asaftei, 2008; Hughes and Mester, 2013; Wheelock and Wilson, 2020).

More concretely, following McCord and Prescott (2014) and the FFIEC 041 Reports, we compute Y_3 by summing credit-equivalent amounts of all off-balance sheet items. For instance, in 2015–2018, Call Reports define these items as off-balance sheet securitization exposures, financial standby letters of credit, performance standby letters of credit and transaction-related contingent items, commercial and similar letters of credit with an original maturity of one year or less, retained recourse on small business obligations sold with recourse, repo-style transactions, unused commitments excluding unused commitments to asset backed commercial paper conduits, unconditionally cancelable commitments, over-the-counter derivatives, centrally cleared derivatives, and all other off-balance sheet liabilities.

We opt for the credit equivalent of off-balance sheet activities over another popular alternative proxy for banks’ nontraditional operations based on net non-interest income (e.g., DeYoung and Rice, 2004; DeYoung and Torna, 2013; Lozano-Vivas and Pasiouras, 2010; Davies and Tracey, 2014; Malikov et al., 2015; Wheelock and Wilson, 2012, 2018) because the latter can be negative, which makes it an undesirable measure for one of the bank’s outputs (see Hughes and Mester, 1998). It is, perhaps, even less suitable a measure for an “output” in the structural production analysis because of its fundamental conceptual incongruity with how the bank’s other outputs are measured following the convention in the literature: namely, it is based on a “flow” (income) data whereas loans Y_1 and securities Y_2 are the “stock” (asset) measures. No such issue arises when using credit equivalents of off-balance sheet items. Having said that, we also redo our analysis using this alternative income-based measure of nontraditional activities in one of the robustness checks. In this case, following the literature, the Y_3 variable is measured using the total non-interest income (inclusive of the income from fiduciary activities, securities brokerage, investment banking, insurance activities, venture capital and the trading revenue) minus service charges on deposit accounts.

The three variable inputs are X_1 — physical capital measured by fixed assets, X_2 — labor, measured as the number of full-time equivalent employees, and X_3 — total borrowed funds, inclusive of deposits and federal funds. Their respective prices are W_1 , W_2 and W_3 , where W_1 is measured as the expenditures on fixed assets divided by premises and fixed assets, W_2 is computed by dividing salaries and employee benefits by the number of full-time equivalent employees, and W_3 is computed as the interest expenses on deposits and fed funds divided by the sum of total deposits and fed funds purchased. Total variable cost C is a sum of expenses on X_1 , X_2 and X_3 .

We also consider equity capital K_1 as an additional input. However, due to the unavailability of the price of equity, we follow Berger and Mester (2003) and Feng and Serletis (2010) in modeling K_1 as a quasi-fixed input. The treatment of equity as an input to banking production technology is consistent with Hughes and Mester (1993, 1998) and Berger and Mester (2003) in that banks may use it as a source of loanable funds and thus as a cushion against losses. By including equity K_1 in the cost analysis, we are therefore also able to control for the bank’s insolvency risk along the lines of Hughes and Mester’s

Table 1. Data Summary Statistics

Variables	Mean	1st Qu.	Median	3rd Qu.
C	13,602.44	2,344.10	4,612.10	10,066.33
Y_1	424,480.49	55,819.58	114,898.89	265,677.61
Y_2	118,621.10	13,161.49	31,803.01	77,360.61
Y_3	27,304.00	659.78	2,847.73	10,503.06
W_1	50.79	14.83	21.27	33.85
W_2	57.86	47.00	54.34	64.90
W_3	0.82	0.40	0.66	1.09
K_1	70,383.27	9,433.85	18,382.47	40,467.85
K_2	0.03	0.01	0.02	0.04
K_3	1.04	0.01	0.08	0.31

C – total variable costs; Y_1 – total loans; Y_2 – total securities; Y_3 – off-balance sheet output measured using credit equivalents; W_1 – price of physical capital; W_2 – price of labor; W_3 – price of financial capital; K_1 – total equity; K_2 – the ratio of nonperforming assets to total assets; K_3 – the ratio of loan loss provisions to total assets. Variables C , W_1 , W_2 , Y_1 , Y_2 , Y_3 , and K_1 are in thousands of real 2005 USD. Variables W_3 , K_2 and K_3 are in %.

(2003) arguments, whereby “an increase in financial capital reduces the probability of insolvency and provides an incentive for allocating additional resources to manage risk in order to protect the larger equity stake” (p.314). In effect, conditioning the bank’s cost on financial capital also allows controlling for quality of loans since the latter is influenced by risk preferences: as Mester (1996) explains, risk-averse bank managers may choose to fund their loans with higher equity-to-deposits ratios (and thus less debt) than a risk-neutral bank would. In our analysis, we also condition the bank’s cost on two other proxy measures of output quality reflective of credit risk associated with the likelihood that borrowers default on their loans and accrued interest by failing to make payments as contractually obligated. We include two most commonly used proxies: the ratio of nonperforming assets to total assets K_2 (e.g., Hughes and Mester, 2013; Wheelock and Wilson, 2018, 2020) and the ratio of loan loss provision to total assets K_3 (e.g., see Laeven and Majnoni, 2003; Acharya et al., 2006; Berger et al., 2010).⁶ analogous to K_1 . Obviously, banks’ expectations of credit risk are unobservable but as noted by Berger et al. (2010), while the former proxy is an *ex-post* measure of the actual incurred losses from lending, the loan loss provisions can be interpreted as an *ex-ante* measure of the level of expected losses and thus as a proxy for expected quality of assets. Controlling for both when modeling bank costs is imperative because lower-quality assets generally require more resources to manage a higher-level risk exposure thereby raising the costs for banks (see Hughes and Mester, 2013). Following the literature, we define nonperforming assets as a sum of total loans and lease financing receivables past due 30 days or more and still accruing, total loans and lease financing receivables not accruing, other real estate owned, and charge-offs on past-due loans and leases. The loss provision is measured using the total provision for loan and lease losses.

We exclude observations that have negative/missing values for assets, equity, output quantities and input prices, which are likely the result of erroneous data reporting. This leaves us with an operational sample of 44,704 observations for 7,232 banks. We deflate all nominal variables to the 2005 U.S. dollars

⁶Although we denote these variables as “ K ,” we do *not* conceptualize them as the quasi-fixed input quantities

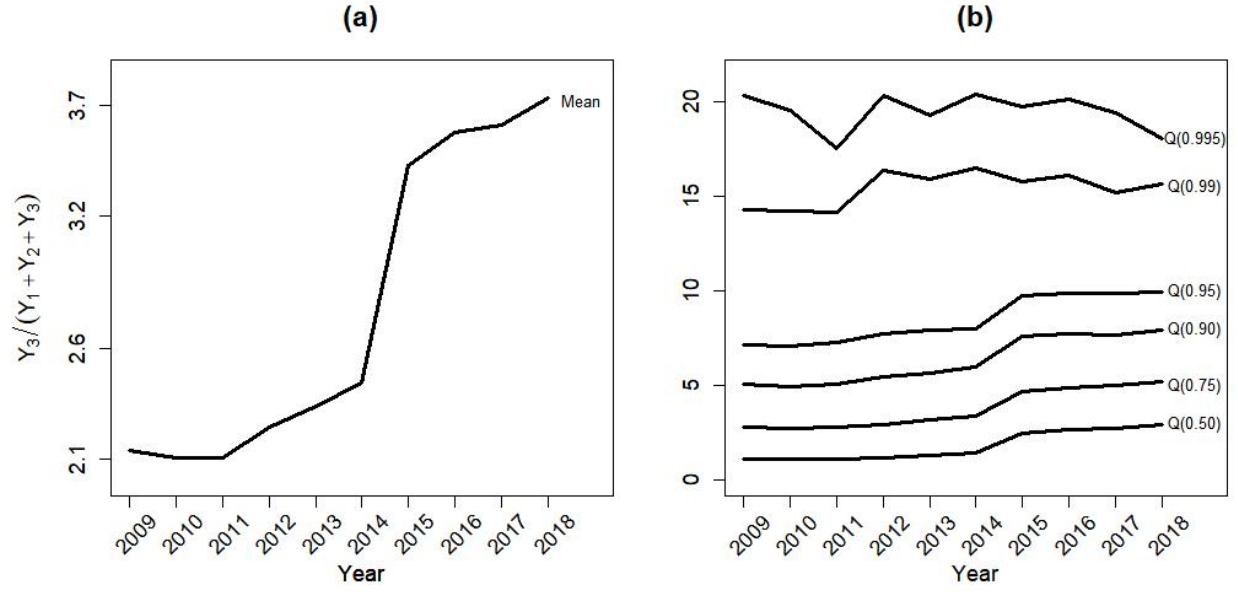


Figure 1. Output Share of Off-Balance sheet Activities over Time, in %: (a) average, (b) select quantiles

using the consumer price index. Table 1 provides summary statistics for our main variables.

Given our emphasis on accounting for banks' off-balance sheet operations, of particular interest is the nontraditional output. Although descriptive statistics in Table 1 expectedly indicate that the volume of Y_3 is significantly smaller than that of the two other traditional outputs, banks' involvement in these off-balance sheet activities has, in fact, been steadily expanding in recent years. To show this, we plot the average share of off-balance sheet activities in bank's total output $Y_3 / (Y_1 + Y_2 + Y_3)$ in Figure 1(a), from where it is evident that the average share of nontraditional outputs among U.S. commercial banks has been steadily increasing since 2011, almost doubling from about 2% in 2009 to 3.7% in 2018. This is consistent with the narrative that commercial banks in the U.S. are increasingly shifting towards off-balance sheet banking.

Obviously, the level of involvement in such nontraditional activities varies considerably across banks, and the rather modest *average* share of the off-balance sheet activities plotted in Figure 1(a) does not provide a complete picture of the growing prevalence of nontraditional activities in banks' operations because it conceals the well-documented heterogeneity across individual banks. For instance, some banks in our sample are highly specialized in off-balance sheet activities, which account for about 70% of their outputs. Therefore, we also examine an evolution of the off-balance sheet share in banks' output portfolio at different quantiles in the data, with the particular focus on the upper tail of the distribution.

Figure 1(b) plots select upper quantiles of $Y_3 / (Y_1 + Y_2 + Y_3)$ over the years, with the lines from bottom to top corresponding to the median, 0.75th, 0.90th, 0.99th and 0.995th quantiles. Two observations are in order here. First, owing to the positive skew in the off-balance sheet share distribution, the differences in banks' involvement in nontraditional banking are stark, with the output share ranging from about 1.8% at the median to 19.5% for banks at the top 0.995th quantile. This cross-bank heterogeneity

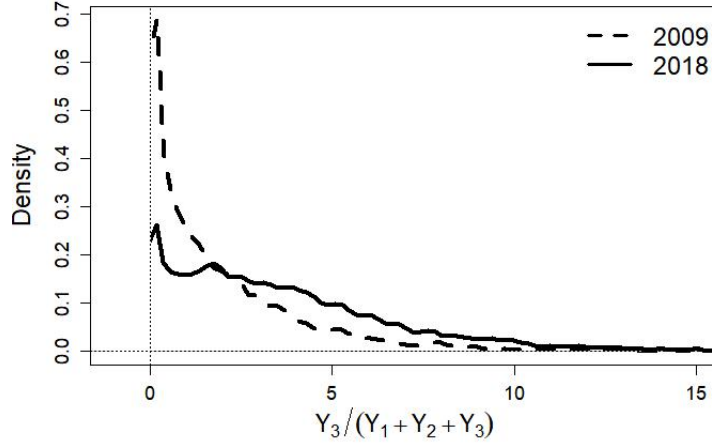


Figure 2. Distribution of the Output Share of Off-Balance Sheet Activities in 2009 vs. 2018

is expected as the choice to engage in nontraditional operations is associated with various idiosyncratic characteristics of banks, including their asset size (Rogers and Sinkey Jr, 1999). Second, the rising share of off-balance sheet activities is present across just about its entire distribution. The latter is particularly evident in Figure 2 that plots this distribution at the beginning (2009) and the end (2018) of our sample period. Altogether, these data document the rising and heterogeneous level of involvement in nontraditional off-balance sheet activities by U.S. banks in the post-crisis period, not only corroborating the common argument that off-balance sheet activities ought to be accounted in the analysis of banks but also illustrating the importance of adequately accommodating vast heterogeneity across banks in that analysis. We seek to address both these imperatives in our paper.

5 Empirical Results

This section reports the results based on our time-varying-coefficient fixed-effects quantile model of bank cost that explicitly accommodates three-way heterogeneity across banks: *(i)* distributional heterogeneity, *(ii)* cross-time heterogeneity and *(iii)* unobserved bank heterogeneity.

Although our analysis is at different quantiles of the bank's *cost*, the interpretation of distribution heterogeneity can be generalized and extended to bank *size* because the bank's operation cost is a good proxy for its size/scale. To sufficiently capture distributional heterogeneity across banks, we estimate our model for the 0.10th, 0.25th, 0.50th, 0.75th and 0.90th quantiles. The middle three quantiles shed light on the cost structure of mid-size banks in the interquartile range of the conditional log-cost distribution, whereas the more extreme 0.10th and 0.90th quantiles provide evidence for the smaller and larger banks, respectively.

For inference, we use the 95% bias-corrected bootstrap percentile confidence intervals: one- or two-sided, as appropriate. In what follows, we discuss our main empirical results pertaining to scope

economies. We then supplement that discussion by also considering two other sources of potential cost savings in banking, namely, scale economies and technological progress.

5.1 Scope Economies

As discussed in Section 2, we investigate the presence of scope economies by using the expansion-path measure of cost subadditivity. Since we analyze bank cost structure across the entire cost distribution as opposed to its first moment (i.e., conditional mean), our cost subadditivity measure is not only observation- but also cost-quantile-specific. When evaluating the formulae in (2.4)–(2.5), we replace $\mathcal{C}_t(\cdot)$ with the exponentiated quantile function of the log-cost $\mathcal{Q}_c(\tau|\cdot)$ since our cost function estimation is for a conditional log-quantile. That is, for a given quantile τ , we compute the cost subadditivity measure as

$$\mathcal{S}_t(\tau) = \frac{\sum_{\kappa} \exp \left[\mathcal{Q}_c \left(\tau | \omega_1^{\kappa} Y_1^* + \min\{Y_1\}, \omega_2^{\kappa} Y_2^* + \min\{Y_2\}, \omega_3^{\kappa} Y_3^* + \min\{Y_3\}, t \right) \right] - \exp \left[\mathcal{Q}_c(\tau | Y_1, Y_2, Y_3, t) \right]}{\exp \left[\mathcal{Q}_c(\tau | Y_1, Y_2, Y_3, t) \right]}. \quad (5.1)$$

It is noteworthy that our use of quantiles offers another advantage over the more traditional conditional-mean models whereby, owing to a “monotone equivariance property” of quantiles, our estimates of $\mathcal{S}_t(\tau)$, which are based on the *level* of cost, are immune to transformation biases due to exponentiation of the estimated *log*-cost function. The same however cannot be said about the estimates of scope economies in analogous conditional-mean analyses. Specifically, to evaluate scope economies, most studies typically exponentiate the predicted *logarithm* of bank cost from the estimated translog conditional-mean regressions while ignoring Jensen’s inequality. Consequently, their scope economies estimates are likely biased. To see this, let the conventional fixed-coefficient translog cost regression be $c = f(\mathbf{v}) + \epsilon$ with $\mathbb{E}[\epsilon|\mathbf{v}] = 0$, and recall that upper/lower-case variables are in levels/logs. It then trivially follows that $\mathbb{E}[C|\mathbf{v}] = \exp\{f(\mathbf{v})\}\mathbb{E}[\exp\{\epsilon\}|\mathbf{v}]$ which generally diverges from $\exp\{f(\mathbf{v})\}$ by a multiplicative function of $\mathbf{v}$. Since cost counterfactuals in $\mathcal{S}_t(\tau)$ admit different “ $\mathbf{v}$ ” values as arguments, the cost subadditivity measure above will normally be biased and need not have the same magnitude or even sign as the true quantity unless $\exp\{\epsilon\}$ is mean-independent of $\mathbf{v}$ which is unlikely to be true in practice, say, if ϵ is heteroskedastic. In the case of quantile estimation, we however do *not* face such a problem owing to the equivariance of quantiles to monotone transformations, viz. $\mathcal{Q}_C[\tau|\mathbf{v}] = \mathcal{Q}_{\exp\{\epsilon\}}[\tau|\mathbf{v}] = \exp\{\mathcal{Q}_c[\tau|\mathbf{v}]\}$ (e.g., see Koenker, 2005).

Now, recall that $\mathcal{S}_t(\tau)$ depends on the choice of $\{\omega_m^{\kappa}\}$, which we circumvent by choosing weights that yield the smallest cost subadditivity measure for a given cost quantile τ in the admissible region: $\mathcal{S}_t^*(\tau)$. Namely, for each fixed cost quantile of interest, we perform a grid search over a permissible range of weights in $[0, 1]^6$ at the 0.1 increments. We do this for each bank in a given year. Table 2 summarizes such point estimates of $\mathcal{S}_t^*(\tau)$ for different quantiles of the conditional cost distribution. (We caution readers against confusing quantiles of the conditional cost distribution τ , for which our bank cost function and the cost subadditivity measure are estimated, with the quantiles of empirical distribution of observation-specific $\mathcal{S}_t^*(\tau)$ estimates corresponding to a given τ .)

The two hypotheses of particular interest here are (i) $\mathbb{H}_0 : \mathcal{S}_t^*(\tau) \leq 0$ v. $\mathbb{H}_1 : \mathcal{S}_t^*(\tau) > 0$ and (ii) $\mathbb{H}_0 :$

Table 2. Cost Subadditivity Estimates

Cost Quantiles (τ)	Mean	Point Estimates			Inference Categories			
		1st Qu.	Median	3rd Qu.	= 0	$\neq 0$	> 0	≤ 0
$\mathcal{Q}(0.10)$	0.138 (0.058, 0.469)	0.078 (0.023, 0.288)	0.125 (0.048, 0.463)	0.181 (0.082, 0.626)	9.76%	90.24%	92.04%	7.96%
$\mathcal{Q}(0.25)$	0.175 (0.078, 0.598)	0.107 (0.036, 0.361)	0.163 (0.067, 0.579)	0.225 (0.106, 0.777)	5.48%	94.52%	95.70%	4.30%
$\mathcal{Q}(0.50)$	0.264 (0.120, 0.937)	0.175 (0.066, 0.549)	0.258 (0.109, 0.873)	0.335 (0.155, 1.185)	1.40%	98.60%	98.90%	1.10%
$\mathcal{Q}(0.75)$	0.388 (0.194, 1.205)	0.259 (0.103, 0.683)	0.394 (0.169, 1.113)	0.496 (0.242, 1.582)	0.45%	99.55%	99.50%	0.50%
$\mathcal{Q}(0.90)$	0.459 (0.261, 1.164)	0.313 (0.121, 0.671)	0.476 (0.231, 1.036)	0.575 (0.356, 1.567)	0.30%	99.70%	99.60%	0.40%

The left panel summarizes point estimates of $\mathcal{S}_t^*(\tau)$ with the corresponding two-sided 95% bias-corrected confidence intervals in parentheses. Each bank-year is classified as exhibiting scope economies [$\mathcal{S}_t^*(\tau) > 0$] vs. non-economies [$\mathcal{S}_t^*(\tau) \leq 0$] and scope invariance [$\mathcal{S}_t^*(\tau) = 0$] vs. scope non-invariance [$\mathcal{S}_t^*(\tau) \neq 0$] using the corresponding one- and two-sided 95% bias-corrected confidence bounds, respectively. The right panel reports sample shares for each category and for its corresponding negating alternative. Percentage points sum up to a hundred within binary groups only.

$\mathcal{S}_t^*(\tau) = 0$ v. $\mathbb{H}_1 : \mathcal{S}_t^*(\tau) \neq 0$. Both tests are essentially the same, except for the one- or two-sided alternatives. Although the (i, t) index on outputs is suppressed in (5.1), the tests are at the level of observation (bank-year). In case of (i), rejection of the null would imply that even the smallest subadditivity measure is statistically *positive* and scope economies can thus be inferred to also be locally significant over the bank's output space in a given year. In case of (ii), failure to reject the null would suggest that subadditivity measure is statistically indistinguishable from zero, which is consistent with the bank's cost structure exhibiting local scope invariance.

The right panel of Table 2 reports the results of these hypothesis tests. Namely, for each cost quantile τ , we classify banks in our data based on the two dichotomous groups of categories: banks that exhibit scope economies [$\mathcal{S}_t^*(\tau) > 0$] vs. scope non-economies [$\mathcal{S}_t^*(\tau) \leq 0$] and the banks whose cost structure that exhibits scope invariance [$\mathcal{S}_t^*(\tau) = 0$] vs. scope non-invariance [$\mathcal{S}_t^*(\tau) \neq 0$].

Our results provide strong evidence in support of statistically significant scope economies across banks virtually of all sizes in the U.S. banking sector. For banks in the middle interquartile range of the cost—essentially, size—distribution, at least 95.7% exhibit positive economies of scope. For the top half of the distribution (median or higher), the prevalence of significant scope economies is about 99%. Even at the very bottom of cost distribution ($\tau = 0.1$) where the revenue diversification opportunities may not be as abundant or easily accessible, our test results suggest that roughly 92% of banks enjoy scope-driven cost savings and those, who do not, exhibit scope invariance. Figure 3 provides a graphic illustration of these results. For each considered cost quantile τ , the figure shows a scatter-plot of the $\mathcal{S}_t^*(\tau)$ estimates for each bank-year observation along with the corresponding one-sided 95% lower confidence bound. Here, we sort these estimates by their lower confidence bounds (solid line) and color them based on whether they are significantly above 0 or not. From Figure 3, it is evident that positive scope economies are ubiquitous and that their presence is only growing with quantiles of the conditional variable-cost distribution of banks (i.e., with the bank size).

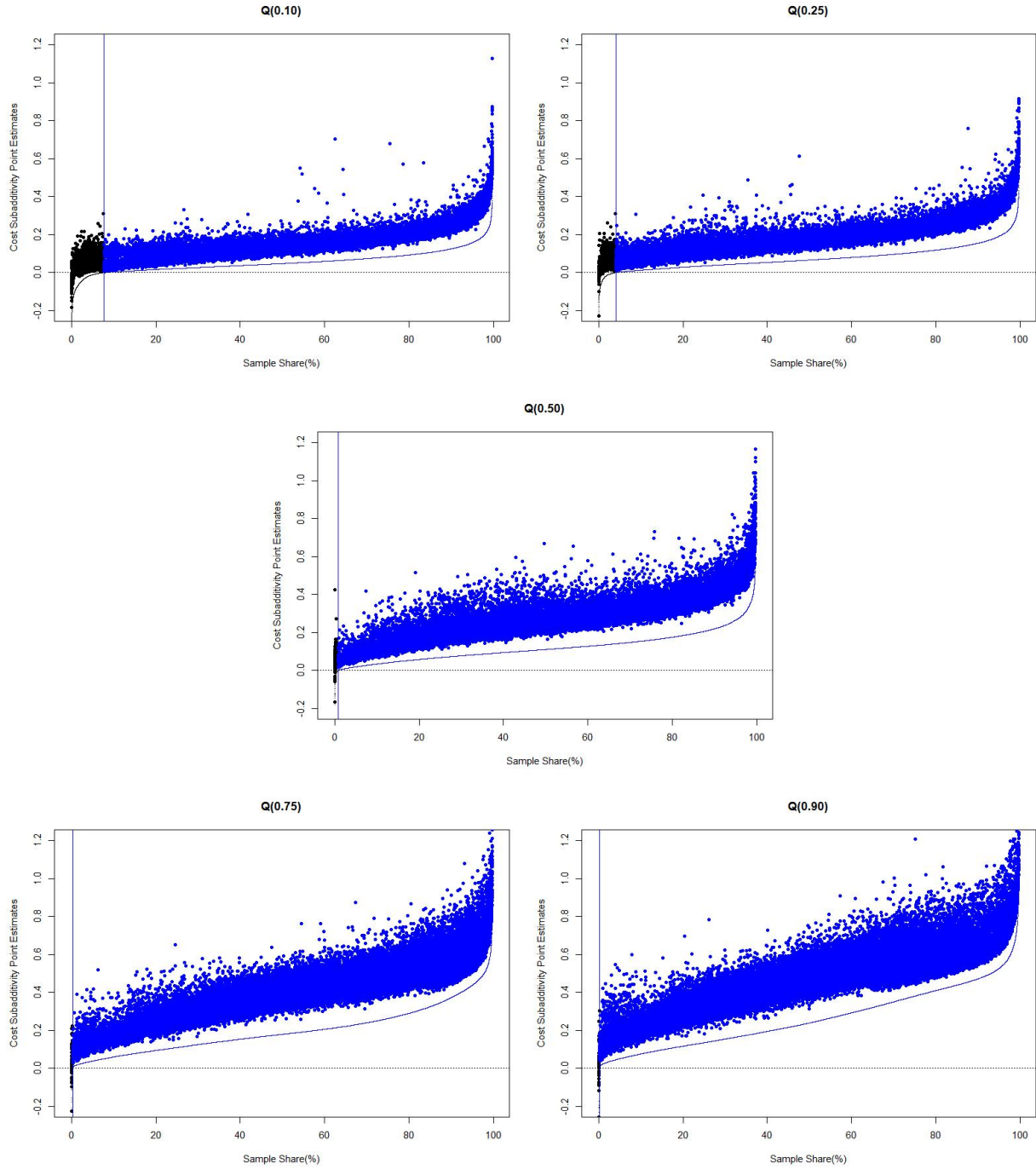


Figure 3. The One-Sided 95% Lower Bounds (solid lines) of the Cost Subadditivity Point Estimates (scatter points) Across Cost Quantiles

Table 3. Cost Subadditivity Estimates: Robustness to Alternative Variable Specifications

	(I): Main Specification			(II)			(III)			(IV)		
Cost Quantiles (τ)	Median Est.	Categories = 0 > 0		Median Est.	Categories = 0 > 0		Median Est.	Categories = 0 > 0		Median Est.	Categories = 0 > 0	
$\mathcal{Q}(0.10)$	0.125	9.8%	92.0%	0.182	2.1%	98.3%	0.068	58.9%	40.8%	0.239	38.0%	65.0%
$\mathcal{Q}(0.25)$	0.163	5.5%	95.7%	0.256	0.6%	99.4%	0.120	41.6%	59.1%	0.282	20.3%	82.3%
$\mathcal{Q}(0.50)$	0.258	1.4%	98.9%	0.429	0.1%	99.7%	0.226	11.4%	89.0%	0.346	3.3%	97.5%
$\mathcal{Q}(0.75)$	0.394	0.5%	99.5%	0.543	0.1%	99.7%	0.356	2.6%	97.6%	0.409	1.0%	99.2%
$\mathcal{Q}(0.90)$	0.476	0.3%	99.6%	0.589	0.0%	99.7%	0.427	1.7%	98.5%	0.447	0.8%	99.4%
Nontraditional Output Measure:												
Credit Equivalents		✓			✓							
Net Non-Interest Income								✓			✓	
Credit Risk Proxies:												
Nonperforming Assets		✓			✓			✓			✓	
Loan Loss Provision		✓						✓				

Reported are the median point estimates of $\mathcal{S}_t^*(\tau)$ and shares of the sample for which the estimates are statistically > 0 (i.e., a bank-year is classified as exhibiting scope economies) and statistically not different from 0 (i.e., a bank-year is classified as exhibiting scope invariance) at the 95% level. Because the two hypotheses are tested separately, percentage points need not sum up to a hundred. Specification (I) is our main specification, the complete results for which are reported in Table 2.

As a robustness check, we re-estimate our model under alternative empirical specifications of the cost-function variables. Namely, we consider a different proxy for nontraditional operations used in the literature (net non-interest income) as well as assess sensitivity of our findings to credit risk proxies included in the analysis. Table 3 summarizes estimates of cost subadditivity across these alternatives. Two observations are in order here. First, omitting an *ex-ante* proxy for output quality (loan loss provisions) produces uniformly larger point estimates of cost subadditivity. Consequently, the evidence in favor of significantly positive scope economies is even stronger in the latter case. Nonetheless, we continue to include this important control in our main specification. Second, when using net non-interest income as a proxy measure of nontraditional banking operations, we obtain smaller $\mathcal{S}_t^*(\tau)$ estimates, with the largest differences seen at the bottom tail of costs. But even then, the empirical evidence in support of scope economies across banks is strong. For banks in the middle of the cost distribution (at the conditional median), 89% exhibit significant economies of scope. The prevalence of product-scope-driven cost savings is even more pervasive ($\geq 97.6\%$) for larger banks at higher quantiles. In the case of smaller-scale banks at the bottom 0.10th and 0.25th quantiles, the share of banks that enjoy scope economies—while smaller—is nonetheless non-negligible, ranging between 41–65% and 59–82%, respectively, depending on the credit risk proxies included in the analysis. The cost structure of the remaining banks is scope-invariant. All in all, our findings of significant scope economies in banking are robust to alternative specifications, and in what follows, we therefore focus on the results from our main specification only.

A finding worth emphasizing here is that, having accounted for three-way heterogeneity across banks in a pursuit of robust estimates of bank cost subadditivity, we find *no* empirical evidence in support of scope *diseconomies*. This is in stark contrast with earlier studies of scope economies in U.S. banking (e.g., Berger et al., 1987; Mester, 1987; Hughes and Mester, 1993; Pulley and Braunstein, 1992; Ferrier et al., 1993; Pulley and Humphrey, 1993; Jagtiani et al., 1995; Jagtiani and Khanthavit, 1996; Wheelock

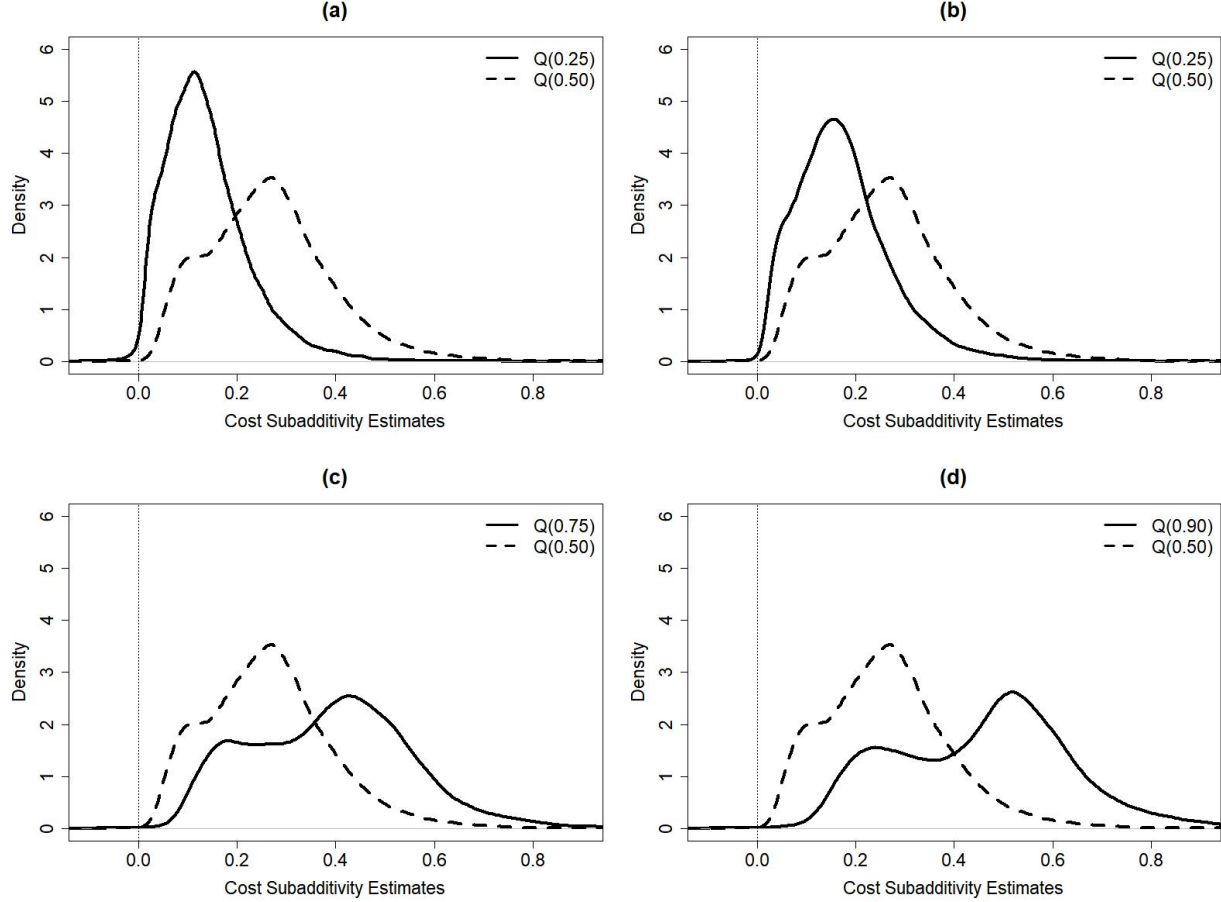


Figure 4. Kernel Densities of Cost Subadditivity Estimates Across Cost Quantiles

and Wilson, 2001). Besides our reliance on the more robust estimation methodology, the qualitative differences between our and prior findings can also be attributed to fundamental changes that the banking sector has undergone in the past two decades characterized by the growing importance of nontraditional banking operations propelled by the financial product innovations.

Although, the subadditivity measure does not directly quantify the *magnitude* of scope economies in the conventional interpretation of the latter, the value of its point estimates can still provide useful insights into the diversification-driven cost savings. Recall that $\mathcal{S}_t^*(\tau)$ compares the cumulative cost of multiple smaller banks of higher degrees of *relative* output specialization with the cost of a larger, more relatively diversified bank. Essentially, the subadditivity measure sheds light on scope economies from a perspective of relative—as opposed to absolute—notation of revenue diversification. Measured is the reduction in bank cost (in proportions) afforded by achieving lower specialization in any one output. From the left panel of Table 2, the mean estimates of cost subadditivity ranges from 0.138 to 0.459 depending on the conditional cost quantile. This suggests, on average, the potential for a 14–46% cost saving if the bank “rebalances” its joint production of loans, securities and off-balance sheet outputs. We also find that the magnitude of diversification-driven economies increases as one moves from the bottom to top

Table 4. Stochastic Dominance of Scope Subadditivity Across Cost Quantiles

	$\{\mathcal{Q}(0.75), \dots, \mathcal{Q}(0.10)\}$	$\{\mathcal{Q}(0.50), \mathcal{Q}(0.25), \mathcal{Q}(0.10)\}$	$\{\mathcal{Q}(0.25), \mathcal{Q}(0.10)\}$	$\mathcal{Q}(0.10)$
$\mathcal{Q}(90)$	0.578	0.578	0.739	0.970
$\mathcal{Q}(75)$		0.894	0.970	0.970
$\mathcal{Q}(50)$			0.970	0.970
$\mathcal{Q}(25)$				0.784
Reported are the p -values.				

of the bank cost distribution, thereby suggesting that larger banks (higher τ) may economize cost better compared to those of smaller size in the lower end of the cost distribution.

For a more holistic look at the empirical evidence of scope economies across different quantiles of the bank cost distribution, we also provide kernel density plots of the $\mathcal{S}_t^*(\tau)$ estimates in Figure 4. It enables us to compare distributions of the cost subadditivity estimates as opposed to merely focusing on marginal moments. Consistent with our earlier discussion, these plots indicate that large-scale banks lying in the upper quantiles of the cost distribution appear to enjoy bigger diversification-driven cost economies than those in the lower cost quantiles. To support this visual evidence, we formally test for the (first-order) stochastic dominance of scope economies exhibited by banks in the top cost quantiles over those exhibited by those in the bottom quantiles. We utilize a generalized Kolmogorov-Smirnov test proposed by Linton et al. (2005) which permits testing dominance over multiple variables (in our case, more than two cost quantiles) and allows these variables to be estimated latent quantities as opposed to observables from the data and to also share dependence (in our case, the dependence is due to common parameter estimates used to construct quantile coefficients). Specifically, let $F_\tau(\mathcal{S})$ represent the cumulative distribution functions of the $\mathcal{S}_t^*(\tau)$ estimates for a given cost quantile τ . We then form the null hypotheses that diversification-driven scope economies exhibited by banks in the lower quantiles of the cost distribution are stochastically dominated by those in the upper quantiles of the cost distribution. More formally, for any cost quantile of interest $\bar{\tau} \in \mathbb{T}$ with $\mathbb{T} = \{0.10, 0.25, 0.50, 0.75, 0.90\}$, we are interested in

$$\mathbb{H}_0 : \min_{\tau \neq \bar{\tau} \in \mathbb{T}} \sup_{\mathcal{S} \in \mathbb{S}} [F_\tau(\mathcal{S}) - F_{\bar{\tau}}(\mathcal{S})] \leq 0 \text{ v. } \mathbb{H}_1 : \min_{\tau \neq \bar{\tau} \in \mathbb{T}} \sup_{\mathcal{S} \in \mathbb{S}} [F_\tau(\mathcal{S}) - F_{\bar{\tau}}(\mathcal{S})] > 0.$$

We use the sub-sampling procedure suggested by Linton et al. (2005) to perform the test.⁷

Table 4 reports p -values for the tests of dominance of $\mathcal{S}_t^*(\tau)$ from the “row” quantile over a multi-quantile set of $\mathcal{S}_t^*(\tau)$ from the “column” quantiles. All p -values are safely greater than the conventional 0.05 level, and we fail to reject the nulls. Combined with the visual evidence from Figure 4, we can therefore infer that bigger banks in the higher quantiles of the cost distribution exhibit larger scope economies than do smaller banks from the lower cost quantiles for the entire set of observable output mixes. Relatedly, of interest is the relation between the scope economies magnitude and the degree of bank’s specialization in nontraditional products. To examine this, for each cost quantile τ that we

⁷We employ 199 equidistant sub-sample sizes $B_n = \{b_1, \dots, b_{199}\}$, where $b_1 = \lfloor \log \log N \rfloor$, $b_{199} = \lfloor N / \log \log N \rfloor$ with $N = nT$ being the sample size. For each sub-sample size, we get a p -value. The reported is the mean of these 199 p -values.

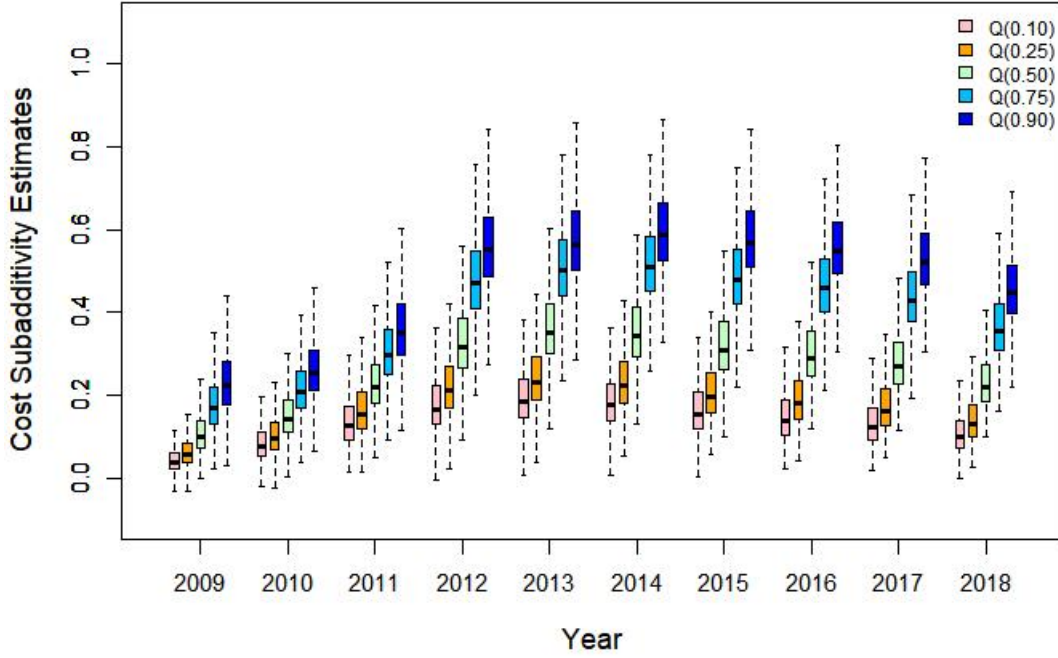


Figure 5. Evolution of Cost Subadditivity

consider in our analysis, we run a least-absolute-deviation regression of the $\mathcal{S}_t^*(\tau)$ estimates on the share of off-balance sheet activities in bank's total output $Y_3/(Y_1 + Y_2 + Y_3)$. Their median *associations* are all significant and monotonically increasing with cost quantile: -0.31 , -0.29 , -0.09 , 0.35 and 0.41 for $\tau = 0.10, 0.25, 0.50, 0.75, 0.90$, respectively. This suggest that, among larger banks (higher τ) those who engage in off-balance sheet banking more heavily tend to enjoy scope economies of greater degrees. In contrast, for smaller banks (lower τ), pivoting off balance sheet is associated with reduced scope-driven cost savings, plausibly because of their limited capabilities to capitalize on cross-output spillovers and input complementarities at smaller operations scales.

Lastly, we take a look at the evolution of scope economies. Figure 5 documents how distributions of the cost subadditivity estimates shifted over time. Plotted are the box-plots of $\mathcal{S}_t^*(\tau)$ across five considered cost quantiles τ for each year t . The data suggest a divergence in the degree of cost subadditivity between smaller (lower cost quantiles) and larger (higher cost quantiles) banks over time which, however, started reverting in the last years of the sample. We further observe that, while positive and significant throughout, the magnitude of a cost-saving potential associated with the product-scope diversification picked around 2013–2014 and has since been in a steady decline, across all cost quantiles.

Table 5. Returns to Scale Estimates

Cost Quantiles (τ)	Mean	Point Estimates			Inference Categories, %			
		1st Qu.	Median	3rd Qu.	= 1	$\neq 1$	> 1	≤ 1
$\mathcal{Q}(0.10)$	1.300 (1.263, 1.352)	1.263 (1.226, 1.316)	1.293 (1.257, 1.344)	1.328 (1.286, 1.382)	0.02	99.98	99.99	0.01
$\mathcal{Q}(0.25)$	1.321 (1.282, 1.363)	1.282 (1.243, 1.322)	1.313 (1.276, 1.354)	1.351 (1.306, 1.399)	0.01	99.99	100.0	0.00
$\mathcal{Q}(0.50)$	1.361 (1.319, 1.404)	1.316 (1.276, 1.356)	1.351 (1.31, 1.393)	1.394 (1.347, 1.444)	0.01	99.99	100.0	0.00
$\mathcal{Q}(0.75)$	1.405 (1.352, 1.457)	1.352 (1.307, 1.398)	1.392 (1.344, 1.443)	1.441 (1.385, 1.500)	0.00	100.0	100.0	0.00
$\mathcal{Q}(0.90)$	1.430 (1.363, 1.491)	1.373 (1.314, 1.421)	1.416 (1.353, 1.469)	1.469 (1.397, 1.533)	0.01	99.99	100.0	0.00

The left panel summarizes point estimates of $\mathcal{R}_t(\tau)$ with the corresponding two-sided 95% bias-corrected confidence intervals in parentheses. Each bank-year is classified as exhibiting IRS [$\mathcal{R}_t(\tau) > 1$] vs. non-IRS [$\mathcal{R}_t(\tau) \leq 1$] and CRS [$\mathcal{R}_t(\tau) = 1$] vs. non-CRS [$\mathcal{R}_t(\tau) \neq 1$] using the corresponding one- and two-sided 95% bias-corrected confidence bounds, respectively. The right panel reports sample shares for each category and for its corresponding negating alternative. Percentage points sum up to a hundred within binary groups only.

5.2 Scale Economies

We complement our analysis of the scope-driven cost savings in the U.S. banking with the examination of economies of scale. Scale economies are said to exist if the banks' average cost declines with equiproportional expansion of its outputs (i.e., with the increase in scale of production). As discussed in the introduction, the latter has been a subject of particular academic interest in face of the post-crisis regulatory reforms in the banking sector.

Our returns to scale measure takes into account quasi-fixity of the equity input per Caves et al. (1981):

$$\mathcal{R}_t(\tau) = (1 - \partial \mathcal{Q}_c(\tau|\cdot) / \partial k_1) / \sum_m \partial \mathcal{Q}_c(\tau|\cdot) / \partial y_m, \quad (5.2)$$

where we replaced the usual $\log \mathcal{C}_t(\cdot)$ with the quantile function of the log-cost $\mathcal{Q}_c(\tau|\cdot)$ in the formula since our cost function estimation is for a conditional quantile. The measure of returns to scale is therefore both observation- and cost-quantile-specific.

Just like in the case of scope economies, for a given τ , we are mainly interested in the following two hypotheses: (i) $\mathbb{H}_0 : \mathcal{R}_t(\tau) \leq 1$ v. $\mathbb{H}_1 : \mathcal{R}_t(\tau) > 1$ and (ii) $\mathbb{H}_0 : \mathcal{R}_t(\tau) = 1$ v. $\mathbb{H}_1 : \mathcal{R}_t(\tau) \neq 1$. In case of (i), rejection of the null would imply that the returns to scale statistically *exceed* 1 implying increasing returns (IRS) and, thus, significant scale economies. In case of (ii), failure to reject the null would suggest that returns to scale are statistically indistinguishable from 1, which is consistent with the bank exhibiting constant returns to scale (CRS) and, hence, scale invariance of costs.

Table 5 summarizes point estimates of the returns to scale for all estimated quantiles of the conditional cost distribution of banks. The right panel of the table reports the results of the hypothesis tests. Namely, reported is the breakdown of banks that exhibit IRS (scale economies) vs. non-IRS (scale non-economies) and of banks that exhibit CRS (scale invariance) vs. non-CRS (scale non-invariance).

The results in Table 5 provide overwhelming evidence of ubiquitous scale economies in the banking sector, across all cost quantiles. The average point estimates of returns to scale ranges from 1.30 to 1.43,

with banks from the higher quantiles of cost distribution exhibiting increasing returns to scale of larger magnitudes compared to those from the lower quantiles. We find that almost every single bank in our sample exhibits statistically significant scale economies (IRS). These results suggest that, when the bank radially expands the scale of its operation, its average variable cost decreases. These findings are consistent with the prior results which however are almost exclusively based on the analyses of bank costs at the conditional *mean* (e.g., Wheelock and Wilson, 2012; Hughes and Mester, 2013; Restrepo-Tobón and Kumbhakar, 2015; Malikov et al., 2015; Restrepo-Tobón et al., 2015; Wheelock and Wilson, 2018). Given that we find evidence of significant scale economies along the entire cost *distribution*, our results provide the robust assurance to these earlier findings reported in the literature.

5.3 Technological Change

We conclude our analysis of bank cost structure by examining temporal shifts in the bank cost frontier in face of technological advancements as well as regulatory changes in the industry in aftermath of the 2008 financial crisis. A cost-diminishing technological change can provide another means for cost savings.

Because we model temporal variation in the cost relationship using discretized time indices, we replace the standard continuous measure of technical change with a discrete dual measure of technological change at each cost quantile τ . Namely, from (3.3), we have

$$\begin{aligned} -\mathcal{T}\mathcal{C}_t(\tau) &\equiv \mathcal{Q}_c(\tau|\cdot, t) - \mathcal{Q}_c(\tau|\cdot, t-1) \\ &= \Delta L(t) + \Delta S(t)q_\tau + \\ &\quad \left[\beta_3 \Delta L(t) + \gamma_3 \Delta S(t)q_\tau \right]' \mathbf{v}_{it} + \frac{1}{2} \left[\beta_4 \Delta L(t) + \gamma_4 \Delta S(t)q_\tau \right]' \text{vec}(\mathbf{v}_{it} \mathbf{v}_{it}'), \end{aligned} \quad (5.3)$$

where $\Delta L(t) = L(t) - L(t-1)$ and $\Delta S(t) = S(t) - S(t-1)$, with its feasible analogue given by

$$\begin{aligned} -TC_t(\tau) &= (\eta_\kappa + \theta_\kappa q_\tau)D_{\kappa,t} - (\eta_{\kappa-1} + \theta_{\kappa-1} q_\tau)D_{\kappa-1,t-1} + \\ &\quad \left[(\beta_3 \eta_\kappa + \gamma_3 \theta_\kappa q_\tau)D_{\kappa,t} - (\beta_3 \eta_{\kappa-1} + \gamma_3 \theta_{\kappa-1} q_\tau)D_{\kappa-1,t-1} \right]' \mathbf{v}_{it} + \\ &\quad \frac{1}{2} \left[(\beta_4 \eta_\kappa + \gamma_4 \theta_\kappa q_\tau)D_{\kappa,t} - (\beta_4 \eta_{\kappa-1} + \gamma_4 \theta_{\kappa-1} q_\tau)D_{\kappa-1,t-1} \right]' \text{vec}[\mathbf{v}_{it} \mathbf{v}_{it}']. \end{aligned} \quad (5.4)$$

The first line in (5.4) corresponds to Hick-neutral component of technological change, whereas the last two lines represent non-neutral change.

The point estimates of technological change at different cost quantiles are summarized in Table 6. The right panel of the table reports results of a one-sided test of $\mathbb{H}_0 : TC_t(\tau) \leq 0$ v. $\mathbb{H}_1 : TC_t(\tau) > 0$, i.e., a test of whether $TC_t(\tau)$ is statistically positive implying that the bank enjoys technological *progress* and, therefore, a *ceteris paribus* cost diminution over time.

Our data suggest that, in the period following the financial crisis, only larger banks in the upper tail of the cost distribution ($\tau \geq 0.75$) have been benefiting from significant cost-diminishing technological advances: 1.5–2.1% p.a., on average. The share of these banks with statistically positive technological change estimates is between 54 and 62%, with the rest of banks in the upper quartile exhibiting no statistically significant cost diminution. For mid-size banks at the median of the cost distribution, techni-

Table 6. Technical Change Estimates

Cost Quantiles (τ)	Mean	Point Estimates			Inference Categories, %			
		1st Qu.	Median	3rd Qu.	= 0	$\neq 0$	> 0	≤ 0
$\mathcal{Q}(0.10)$	−0.010 (−0.022, 0.004)	−0.026 (−0.039, −0.010)	−0.008 (−0.020, 0.006)	0.006 (−0.007, 0.021)	63.06	36.94	9.69	90.31
$\mathcal{Q}(0.25)$	−0.005 (−0.015, 0.007)	−0.019 (−0.031, −0.007)	−0.003 (−0.013, 0.009)	0.011 (0.000, 0.025)	61.23	38.77	17.52	82.48
$\mathcal{Q}(0.50)$	0.005 (−0.004, 0.015)	−0.008 (−0.018, 0.001)	0.007 (−0.002, 0.018)	0.021 (0.011, 0.037)	53.71	46.29	36.91	63.09
$\mathcal{Q}(0.75)$	0.015 (0.006, 0.026)	0.002 (−0.008, 0.013)	0.017 (0.008, 0.030)	0.031 (0.019, 0.049)	47.22	52.78	54.33	45.67
$\mathcal{Q}(0.90)$	0.021 (0.008, 0.032)	0.007 (−0.005, 0.019)	0.022 (0.010, 0.035)	0.037 (0.022, 0.056)	42.91	57.09	61.50	38.50

The left panel summarizes point estimates of $TC_t(\tau)$ with the corresponding two-sided 95% bias-corrected confidence intervals in parentheses. Each bank-year is classified as exhibiting technical progress [$TC_t(\tau) > 0$] vs. non-progress [$TC_t(\tau) \leq 0$] and technical stasis [$TC_t(\tau) = 0$] vs. non-stasis [$TC_t(\tau) \neq 0$] using the corresponding one- and two-sided 95% bias-corrected confidence bounds, respectively. The right panel reports sample shares for each category and for its corresponding negating alternative. Percentage points sum up to a hundred within binary groups only.

cal change is statistically positive for modest 37% of banks. Evidence of significant cost diminution is even weaker among banks in the bottom half of the cost distribution. Overall, our results suggest that the cost-saving effects of many recent technological advancements in the banking industry, such as the growing networks of automated teller machines, growing credit card networks, electronic payments, internet banking, etc., that were found in the pre-crisis period by earlier studies (e.g., Wheelock and Wilson, 1999; Almanidis, 2013; Malikov et al., 2015) have now largely waned, plausibly because most banks had already capitalized on them to the fullest extent feasible and/or because they now face new regulatory controls. A significant technical change among larger banks in the upper tail of the cost distribution is likely due to their better capability to adapt and innovate.

6 Conclusion

Propelled by the recent financial product innovations, banks are becoming more complex, branching out into many “nontraditional” banking operations beyond issuance of loans. This broadening of operational scope in a pursuit of revenue diversification may be beneficial if banks exhibit scope economies. The existing empirical evidence lends no support for such product-scope-driven cost economies in banking, but it is greatly outdated and, surprisingly, there has been little (if any) research on this subject despite the drastic transformations that the U.S. banking industry has undergone over the past two decades in the wake of technological advancements and regulatory changes. Commercial banks have significantly shifted towards nontraditional operations, and the portfolio of products offered by present-day banks is very different from that two decades ago. This underscore the importance of taking a fresh look at scope economies in banks because leveraging operational scope continues to play a vital role in operations management in banking. It is also important from a policy evaluation perspective, in the face of new financial regulations such as the Dodd–Frank Wall Street Reform and the Consumer Protection

Act of 2010 that seek to set restrictions on the scale and scope of bank operations.

This paper provides new evidence about scope economies in U.S. commercial banking during the 2009–2018 post-crisis period. We improve upon the prior literature not only by analyzing the most recent and relevant data and accounting for bank’s nontraditional off-balance sheet operations, but also in multiple methodological ways as follows. In a pursuit of robust estimates of scope economies and statistical inference thereon, we estimate a flexible, yet parsimonious, time-varying-coefficient panel-data quantile regression model which accommodates three-way bank heterogeneity: (i) distributional heterogeneity in the cost structure of banks along the size of their costs, (ii) temporal variation in cost complementarities and spillovers due to technological change/innovation, and (iii) unobserved bank confounders such as latent management quality. Our results provide strong evidence in support of significantly positive scope economies across banks of virtually all sizes. Contrary to earlier studies, we find no empirical corroboration for scope diseconomies.

References

- Acharya, V. V., Hasan, I., and Saunders, A. (2006). Should banks be diversified? Evidence from individual bank loan portfolios. *Journal of Business*, 79(3):1355–1412.
- Almanidis, P. (2013). Accounting for heterogeneous technologies in the banking industry: A time-varying stochastic frontier model with threshold effects. *Journal of Productivity Analysis*, 39(2):191–205.
- Apergis, N. (2014). The long-term role of non-traditional banking in profitability and risk profiles: Evidence from a panel of US banking institutions. *Journal of International Money and Finance*, 45:61–73.
- Asaftei, G. (2008). The contribution of product mix versus efficiency and technical change in US banking. *Journal of Banking & Finance*, 32(11):2336–2345.
- Baltagi, B. H. and Griffin, J. M. (1988). A general index of technical change. *Journal of Political Economy*, 96(1):20–41.
- Baumol, W., Panzar, J., and Willig, R. (1982). *Contestable Markets and the Theory of Industry Structure*. Harcourt, Brace & Jovanovich, San Diego.
- Berger, A. N., Hanweck, G. A., and Humphrey, D. B. (1987). Competitive viability in banking: Scale, scope, and product mix economies. *Journal of Monetary Economics*, 20(3):501–520.
- Berger, A. N., Hasan, I., and Zhou, M. (2010). The effects of focus versus diversification on bank performance: Evidence from Chinese banks. *Journal of Banking and Finance*, 34(7):1417–1435.
- Berger, A. N. and Mester, L. J. (2003). Explaining the dramatic changes in performance of US banks: Technological change, deregulation, and dynamic changes in competition. *Journal of Financial Intermediation*, 12(1):57–95.
- Berger, A. N. and Udell, G. F. (1995). Relationship lending and lines of credit in small firm finance. *Journal of Business*, pages 351–381.
- Canay, I. A. (2011). A simple approach to quantile regression for panel data. *Econometrics Journal*, 14:368–386.
- Casu, B. and Girardone, C. (2005). An analysis of the relevance of off-balance sheet items in explaining productivity change in European banking. *Applied Financial Economics*, 15(15):1053–1061.
- Caves, D. W., Christensen, L. R., and Swanson, J. A. (1981). Productivity growth, scale economies, and capacity utilization in US railroads, 1955–74. *American Economic Review*, 71(5):994–1002.
- Clark, J. A. and Siems, T. F. (2002). X-efficiency in banking: Looking beyond the balance sheet. *Journal of Money, Credit and Banking*, pages 987–1013.

- Das, S. R. and Nanda, A. (1999). A theory of banking structure. *Journal of Banking and Finance*, 23(6):863–895.
- Davies, R. and Tracey, B. (2014). Too big to be efficient? The impact of implicit subsidies on estimates of scale economies for banks. *Journal of Money, Credit and Banking*, 46:219–253.
- Degryse, H. and Van Cayseele, P. (2000). Relationship lending within a bank-based system: Evidence from european small business data. *Journal of Financial Intermediation*, 9(1):90–109.
- DeYoung, R. and Rice, T. (2004). How do banks make money? The fallacies of fee income. *Federal Reserve Bank of Chicago, Economic Perspectives*, 4Q:34–51.
- DeYoung, R. and Torna, G. (2013). Nontraditional banking activities and bank failures during the financial crisis. *Journal of Financial Intermediation*, 22(3):397–421.
- Efron, B. (1982). *The Jackknife, the Bootstrap, and Other Resampling Plans*. Society for Industrial and Applied Mathematics.
- Evans, D. S. and Heckman, J. J. (1984). A test for subadditivity of the cost function with an application to the Bell System. *American Economic Review*, 74(4):615–623.
- Feng, G. and Serletis, A. (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 and Finance*, 34(1):127–138.
- Ferrier, G. D., Grosskopf, S., Hayes, K. J., and Yaisawarng, S. (1993). Economies of diversification in the banking industry: A frontier approach. *Journal of Monetary Economics*, 31(2):229–249.
- Hassan, M. K. (1993). The off-balance sheet banking risk of large US commercial banks. *The Quarterly Review of Economics and Finance*, 33(1):51–69.
- Hassan, M. K. and Sackley, W. H. (1994). A methodological investigation of risk exposure of bank off-balance sheet loan commitment activities. *The Quarterly Review of Economics and Finance*, 34(3):283–299.
- Hughes, J. P. and Mester, L. J. (1993). A quality and risk-adjusted cost function for banks: Evidence on the “too-big-to-fail” doctrine. *Journal of Productivity Analysis*, 4(3):293–315.
- Hughes, J. P. and Mester, L. J. (1998). Bank capitalization and cost: Evidence of scale economies in risk management and signaling. *Review of Economics and Statistics*, 80(2):314–325.
- Hughes, J. P. and Mester, L. J. (2013). Who said large banks don’t experience scale economies? Evidence from a risk-return-driven cost function. *Journal of Financial Intermediation*, 22(4):559–585.
- Hughes, J. P. and Mester, L. J. (2015). Measuring the performance of banks: Theory, practice, evidence, and some policy implications. In Berger, A., Molyneux, P., and Wilson, J., editors, *Oxford Handbook of Banking*, pages 247–270. Oxford University Press, Oxford, 2 edition.
- Jagtiani, J. and Khanthavit, A. (1996). Scale and scope economies at large banks: Including off-balance sheet products and regulatory effects (1984–1991). *Journal of Banking & Finance*, 20(7):1271–1287.
- Jagtiani, J., Nathan, A., and Sick, G. (1995). Scale economies and cost complementarities in commercial banks: On-and off-balance-sheet activities. *Journal of Banking and Finance*, 19(7):1175–1189.
- Koenker, R. (2004). Quantile regression for longitudinal data. *Journal of Multivariate Analysis*, 91:74–89.
- Koenker, R. (2005). *Quantile Regression*. Cambridge University Press, New York.
- Koenker, R. and Bassett, G. (1982). Robust test for heteroskedasticity based on regression quantiles. *Econometrica*, 50:43–61.
- Koetter, M., Kolari, J. W., and Spierdijk, L. (2012). Enjoying the quiet life under deregulation? Evidence from adjusted Lerner indices for U.S. banks. *Review of Economics and Statistics*, 94:462–480.
- Laeven, L. and Levine, R. (2007). Is there a diversification discount in financial conglomerates? *Journal of Financial Economics*, 85(2):331–367.
- Laeven, L. and Majnoni, G. (2003). Loan loss provisioning and economic slowdowns: too much , too late? *Journal*

- of Financial Intermediation*, 12:178–197.
- Lamarche, C. (2010). Robust penalised quantile regression estimation for panel data. *Journal of Econometrics*, 157:396–408.
- Linton, O., Maasoumi, E., and Whang, Y.-J. (2005). Consistent testing for stochastic dominance under general sampling schemes. *Review of Economic Studies*, 72(3):735–765.
- Lozano-Vivas, A. and Pasiouras, F. (2010). The impact of non-traditional activities on the estimation of bank efficiency: International evidence. *Journal of Banking and Finance*, 34(7):1436–1449.
- Machado, J. A. F. and Santos Silva, J. M. C. (2019). Quantiles via moments. *Journal of Econometrics*, 213(1):145–173.
- Malikov, E., Restrepo-Tobón, D., and Kumbhakar, S. C. (2015). Estimation of banking technology under credit uncertainty. *Empirical Economics*, 49(1):185–211.
- Markides, C. and Williamson, P. J. (1994). Related diversification, core competences and corporate performance. *Strategic Management Journal*, 15(S2):149–165.
- McCord, R. and Prescott, E. S. (2014). The financial crisis, the collapse of bank entry, and changes in the size distribution of banks. *FRB Richmond Economic Quarterly*, 100(1):23–50.
- Mester, L. J. (1987). A multiproduct cost study of savings and loans. *Journal of Finance*, 42(2):423–445.
- Mester, L. J. (1992). Traditional and nontraditional banking: An information-theoretic approach. *Journal of Banking and Finance*, 16(3):545–566.
- Mester, L. J. (1996). A study of bank efficiency taking into account risk-preferences. *Journal of Banking and Finance*, 20:1025–1045.
- Milgrom, P. and Roberts, J. (1995). Complementarities and fit strategy, structure, and organizational change in manufacturing. *Journal of Accounting and Economics*, 19(2–3):179–208.
- Panzar, J. C. and Willig, R. D. (1981). Economies of scope. *American Economic Review*, 71(2):268–272.
- Perera, A., Ralston, D., and Wickramanayake, J. (2014). Impact of off-balance sheet banking on the bank lending channel of monetary transmission: Evidence from South Asia. *Journal of International Financial Markets, Institutions and Money*, 29:195–216.
- Pulley, L. B. and Braunstein, Y. M. (1992). A composite cost function for multiproduct firms with an application to economies of scope in banking. *Review of Economics and Statistics*, pages 221–230.
- Pulley, L. B. and Humphrey, D. B. (1993). The role of fixed costs and cost complementarities in determining scope economies and the cost of narrow banking proposals. *Journal of Business*, pages 437–462.
- Restrepo-Tobón, D. and Kumbhakar, S. C. (2015). Nonparametric estimation of returns to scale using input distance functions: An application to large U.S. banks. *Empirical Economics*, 48:143–168.
- Restrepo-Tobón, D., Kumbhakar, S. C., and Sun, K. (2015). Obelix vs. asterix: Size of US commercial banks and its regulatory challenge. *Journal of Regulatory Economics*, 48:125–168.
- Rime, B. and Stiroh, K. J. (2003). The performance of universal banks: Evidence from Switzerland. *Journal of Banking and Finance*, 27(11):2121–2150.
- Rogers, K. and Sinkey Jr, J. F. (1999). An analysis of nontraditional activities at US commercial banks. *Review of Financial Economics*, 8(1):25–39.
- Rossi, S. P. S., Schwaiger, M. S., and Winkler, G. (2009). How loan portfolio diversification affects risk, efficiency and capitalization: A managerial behavior model for Austrian banks. *Journal of Banking and Finance*, 33:2218–2226.
- Rumelt, R. P. (1982). Diversification strategy and profitability. *Strategic Management Journal*, 3(4):359–369.
- Sealey, C. W. and Lindley, J. T. (1977). Inputs, outputs, and a theory of production and cost at depository financial institutions. *Journal of Finance*, 32(4):1251–1266.
- Skinner, W. (1974a). Decline, fall, and renewal of manufacturing plants. *Industrial Engineering*, 6(10):32–38.
- Skinner, W. (1974b). The focused factory. *Harvard Business Review*, 52:113–121.

- Stiroh, K. J. (2000). How did bank holding companies prosper in the 1990s? *Journal of Banking & Finance*, 24(11):1703–1745.
- Stiroh, K. J. (2004). Diversification in banking: Is noninterest income the answer? *Journal of Money, Credit and Banking*, pages 853–882.
- Villalonga, B. (2004). Diversification discount or premium? New evidence from the Business Information Tracking Series. *Journal of Finance*, 59(2):479–506.
- Wheelock, D. C. and Wilson, P. W. (1999). Technical progress, inefficiency, and productivity change in US banking, 1984-1993. *Journal of Money, Credit, and Banking*, pages 212–234.
- Wheelock, D. C. and Wilson, P. W. (2001). New evidence on returns to scale and product mix among US commercial banks. *Journal of Monetary Economics*, 47(3):653–674.
- Wheelock, D. C. and Wilson, P. W. (2012). Do large banks have lower costs? New estimates of returns to scale for US banks. *Journal of Money, Credit and Banking*, 44(1):171–199.
- Wheelock, D. C. and Wilson, P. W. (2018). The evolution of scale economies in US banking. *Journal of Applied Econometrics*, 33(1):16–28.
- Wheelock, D. C. and Wilson, P. W. (2020). New estimates of the lerner index of market power for U.S. banks. *Federal Reserve Bank of St. Louis Working Paper 2019-012*.
- Yuan, Y. and Phillips, R. D. (2008). Financial integration and scope efficiency in US financial services post Gramm-Leach-Bliley. *Journal of Banking and Finance*.

Appendix

A Three-Step Estimation Procedure

This appendix describes the estimation details of a conditional quantile function in (3.5). First, for ease of notation, we define $\mathbf{D}_t = [D_{2,t}, \dots, D_{T,t}]'$, $\boldsymbol{\eta} = [\eta_2, \dots, \eta_T]'$ and $\boldsymbol{\theta} = [\theta_2, \dots, \theta_T]'$.

Step 1. We first estimate parameters of the location function. Under the assumption (ii), from (3.1) it follows that the conditional mean function of the log-cost c_{it} is

$$\mathbb{E}[c_{it} | \mathbf{v}_{it}, \mathbf{D}_t] = \beta_0 + \sum_{\kappa} \eta_{\kappa} D_{\kappa,t} + \left[\boldsymbol{\beta}_1 + \boldsymbol{\beta}_1^* \sum_{\kappa} \eta_{\kappa} D_{\kappa,t} \right]' \mathbf{v}_{it} + \frac{1}{2} \left[\boldsymbol{\beta}_2 + \boldsymbol{\beta}_2^* \sum_{\kappa} \eta_{\kappa} D_{\kappa,t} \right]' \text{vec}(\mathbf{v}_{it} \mathbf{v}_{it}') + \lambda_i, \quad (\text{A.1})$$

which can be consistently estimated in the within-transformed form via nonlinear least squares after purging additive location fixed effects.⁸

Given the nonlinearity and high dimensionality of (A.1) in parameters, we estimate the slope coefficients via concentration by noticing that, conditional on $\boldsymbol{\eta}$, this mean regression is linear in $[\boldsymbol{\beta}_1', \boldsymbol{\beta}_2', \boldsymbol{\beta}_1^{*'}, \boldsymbol{\beta}_2^{*'}]'$ yielding the profiled least-squares estimator for $[\boldsymbol{\beta}_1(\boldsymbol{\eta})', \boldsymbol{\beta}_2(\boldsymbol{\eta})', \boldsymbol{\beta}_1^*(\boldsymbol{\eta})', \boldsymbol{\beta}_2^*(\boldsymbol{\eta})']'$. Specifically, letting the concentrated sum of (within-transformed) squared errors be

$$\begin{aligned} M(\boldsymbol{\eta}) = \sum_i \sum_t & \left[c_{it} - \bar{c}_i - \boldsymbol{\eta}'(\mathbf{D}_t - \bar{\mathbf{D}}) - (\mathbf{v}_{it} - \bar{\mathbf{v}}_i)' \boldsymbol{\beta}_1(\boldsymbol{\eta}) - \right. \\ & \left. \left(\boldsymbol{\eta}' \mathbf{D}_t \cdot \mathbf{v}_{it} - \overline{\boldsymbol{\eta}' \mathbf{D}_t \cdot \mathbf{v}_i} \right)' \boldsymbol{\beta}_1^*(\boldsymbol{\eta}) - \frac{1}{2} \left(\text{vec}(\mathbf{v}_{it} \mathbf{v}_{it}') - \overline{\text{vec}(\mathbf{v}_i \mathbf{v}_i')} \right)' \boldsymbol{\beta}_2(\boldsymbol{\eta}) - \right. \\ & \left. \frac{1}{2} \left(\boldsymbol{\eta}' \mathbf{D}_t \cdot \text{vec}(\mathbf{v}_{it} \mathbf{v}_{it}') - \overline{\boldsymbol{\eta}' \mathbf{D}_t \cdot \text{vec}(\mathbf{v}_i \mathbf{v}_i')} \right)' \boldsymbol{\beta}_2^*(\boldsymbol{\eta}) \right]^2, \end{aligned} \quad (\text{A.2})$$

with the “bar” denoting the cross-time averages of variables that it tops, we have the profiled estimators $[\boldsymbol{\beta}_1(\boldsymbol{\eta})', \boldsymbol{\beta}_2(\boldsymbol{\eta})', \boldsymbol{\beta}_1^*(\boldsymbol{\eta})', \boldsymbol{\beta}_2^*(\boldsymbol{\eta})']' = (\sum_i \sum_t \mathbb{X}_{it} \mathbb{X}_{it}')^{-1} \sum_i \sum_t \mathbb{X}_{it} \mathbb{Y}_{it}'$, where $\mathbb{X}_{it} = [(\mathbf{v}_{it} - \bar{\mathbf{v}}_i)', (\boldsymbol{\eta}' \mathbf{D}_t \cdot \mathbf{v}_{it} - \overline{\boldsymbol{\eta}' \mathbf{D}_t \cdot \mathbf{v}_i})', \frac{1}{2}(\text{vec}(\mathbf{v}_{it} \mathbf{v}_{it}') - \overline{\text{vec}(\mathbf{v}_i \mathbf{v}_i')})', \frac{1}{2}(\boldsymbol{\eta}' \mathbf{D}_t \cdot \text{vec}(\mathbf{v}_{it} \mathbf{v}_{it}') - \overline{\boldsymbol{\eta}' \mathbf{D}_t \cdot \text{vec}(\mathbf{v}_i \mathbf{v}_i')})']'$ and $\mathbb{Y}_{it}' = c_{it} - \bar{c}_i - \boldsymbol{\eta}'(\mathbf{D}_t - \bar{\mathbf{D}})$.

Thus, the nonlinear fixed-effects estimators of the slope coefficients $[\boldsymbol{\eta}', \boldsymbol{\beta}_1', \boldsymbol{\beta}_2', \boldsymbol{\beta}_1^{*'}, \boldsymbol{\beta}_2^{*'}]'$ in the location functions are

$$\hat{\boldsymbol{\eta}} = \arg \min_{\boldsymbol{\eta}} M(\boldsymbol{\eta}) \quad \text{and} \quad \hat{\boldsymbol{\beta}}_1 = \boldsymbol{\beta}_1(\hat{\boldsymbol{\eta}}), \quad \hat{\boldsymbol{\beta}}_2 = \boldsymbol{\beta}_2(\hat{\boldsymbol{\eta}}), \quad \hat{\boldsymbol{\beta}}_1^* = \boldsymbol{\beta}_1^*(\hat{\boldsymbol{\eta}}), \quad \hat{\boldsymbol{\beta}}_2^* = \boldsymbol{\beta}_2^*(\hat{\boldsymbol{\eta}}). \quad (\text{A.3})$$

Under the usual $\sum_{i=1}^n \lambda_i = 0$ normalization, we can then recover the location-shifting intercept β_0 and fixed effects $\{\lambda_i\}$ via

$$\hat{\beta}_0 = \frac{1}{nT} \sum_i \sum_t \left(c_{it} - \hat{\boldsymbol{\eta}}' \mathbf{D}_t - [\hat{\boldsymbol{\beta}}_1 + \hat{\boldsymbol{\eta}}' \mathbf{D}_t \cdot \hat{\boldsymbol{\beta}}_1^*]' \mathbf{v}_{it} - \frac{1}{2} [\hat{\boldsymbol{\beta}}_2 + \hat{\boldsymbol{\eta}}' \mathbf{D}_t \cdot \hat{\boldsymbol{\beta}}_2^*]' \text{vec}(\mathbf{v}_{it} \mathbf{v}_{it}') \right), \quad (\text{A.4})$$

$$\hat{\lambda}_i = \frac{1}{T} \sum_t \left(c_{it} - \hat{\beta}_0 - \hat{\boldsymbol{\eta}}' \mathbf{D}_t - [\hat{\boldsymbol{\beta}}_1 + \hat{\boldsymbol{\eta}}' \mathbf{D}_t \cdot \hat{\boldsymbol{\beta}}_1^*]' \mathbf{v}_{it} - \frac{1}{2} [\hat{\boldsymbol{\beta}}_2 + \hat{\boldsymbol{\eta}}' \mathbf{D}_t \cdot \hat{\boldsymbol{\beta}}_2^*]' \text{vec}(\mathbf{v}_{it} \mathbf{v}_{it}') \right) \forall i. \quad (\text{A.5})$$

⁸Note that, although (A.1) is nonlinear, the presence of fixed effects does not give rise to the incidental parameter problem in this case because $\{\lambda_i\}$ enters the model additively and is not inside the nonlinear mean function.

Hence, the residual estimator is

$$\hat{u}_{it} = c_{it} - \hat{\beta}_0 - \hat{\eta}' \mathbf{D}_t - \left[\hat{\beta}_1 + \hat{\eta}' \mathbf{D}_t \cdot \hat{\beta}_1^* \right]' \mathbf{v}_{it} - \frac{1}{2} \left[\hat{\beta}_2 + \hat{\eta}' \mathbf{D}_t \cdot \hat{\beta}_2^* \right]' \text{vec}(\mathbf{v}_{it} \mathbf{v}_{it}') - \hat{\lambda}_i. \quad (\text{A.6})$$

Step 2. We then estimate parameters of the scale function. Based on the assumptions (ii)–(iii), we have an auxiliary conditional mean regression:

$$\mathbb{E}[|u_{it}| | \mathbf{v}_{it}, \mathbf{D}_t] = \gamma_0 + \sum_{\kappa} \theta_{\kappa} D_{\kappa,t} + \left[\gamma_1 + \gamma_1^* \sum_{\kappa} \theta_{\kappa} D_{\kappa,t} \right]' \mathbf{v}_{it} + \frac{1}{2} \left[\gamma_2 + \gamma_2^* \sum_{\kappa} \theta_{\kappa} D_{\kappa,t} \right]' \text{vec}(\mathbf{v}_{it} \mathbf{v}_{it}') + \sigma_i, \quad (\text{A.7})$$

which, just like in the first step, we can estimate via nonlinear least squares after within-transforming scale fixed effects out. Concretely, with the concentrated squared residual objective function

$$\begin{aligned} M(\boldsymbol{\theta}) = \sum_i \sum_t & \left[|\hat{u}_{it}| - |\bar{u}_i| - \boldsymbol{\theta}'(\mathbf{D}_t - \bar{\mathbf{D}}) - (\mathbf{v}_{it} - \bar{\mathbf{v}}_i)' \boldsymbol{\gamma}_1(\boldsymbol{\theta}) - \right. \\ & \left. (\boldsymbol{\theta}' \mathbf{D}_t \cdot \mathbf{v}_{it} - \overline{\boldsymbol{\theta}' \mathbf{D}_t \cdot \mathbf{v}_i})' \boldsymbol{\gamma}_1^*(\boldsymbol{\theta}) - \frac{1}{2} \left(\text{vec}(\mathbf{v}_{it} \mathbf{v}_{it}') - \overline{\text{vec}(\mathbf{v}_i \mathbf{v}_i')} \right) \boldsymbol{\gamma}_2(\boldsymbol{\theta}) - \right. \\ & \left. \frac{1}{2} \left(\boldsymbol{\theta}' \mathbf{D}_t \cdot \text{vec}(\mathbf{v}_{it} \mathbf{v}_{it}') - \overline{\boldsymbol{\theta}' \mathbf{D}_t \cdot \text{vec}(\mathbf{v}_i \mathbf{v}_i')} \right)' \boldsymbol{\gamma}_2^*(\boldsymbol{\theta}) \right]^2, \end{aligned} \quad (\text{A.8})$$

and the corresponding profiled estimators given by $[\boldsymbol{\gamma}_1(\boldsymbol{\theta})', \boldsymbol{\gamma}_2(\boldsymbol{\theta})', \boldsymbol{\gamma}_1^*(\boldsymbol{\theta})', \boldsymbol{\gamma}_2^*(\boldsymbol{\theta})']' = (\sum_i \sum_t \mathcal{X}_{it} \mathcal{X}_{it}')^{-1} \times \sum_i \sum_t \mathcal{X}_{it} \mathcal{Y}_{it}'$, where $\mathcal{X}_{it} = [(\mathbf{v}_{it} - \bar{\mathbf{v}}_i)', (\boldsymbol{\theta}' \mathbf{D}_t \cdot \mathbf{v}_{it} - \overline{\boldsymbol{\theta}' \mathbf{D}_t \cdot \mathbf{v}_i})', \frac{1}{2}(\text{vec}(\mathbf{v}_{it} \mathbf{v}_{it}') - \overline{\text{vec}(\mathbf{v}_i \mathbf{v}_i')})', \frac{1}{2}(\boldsymbol{\theta}' \mathbf{D}_t \cdot \text{vec}(\mathbf{v}_{it} \mathbf{v}_{it}') - \overline{\boldsymbol{\theta}' \mathbf{D}_t \cdot \text{vec}(\mathbf{v}_i \mathbf{v}_i')})']'$ and $\mathcal{Y}_{it}' = |\hat{u}_{it}| - |\bar{u}_i| - \boldsymbol{\theta}'(\mathbf{D}_t - \bar{\mathbf{D}})$, the nonlinear fixed-effects estimators of the scale-function slope coefficients $[\boldsymbol{\theta}', \boldsymbol{\gamma}_1', \boldsymbol{\gamma}_1'^*, \boldsymbol{\gamma}_2', \boldsymbol{\gamma}_2'^*]'$ are

$$\hat{\boldsymbol{\theta}} = \arg \min_{\boldsymbol{\theta}} M(\boldsymbol{\theta}) \quad \text{and} \quad \hat{\boldsymbol{\gamma}}_1 = \boldsymbol{\gamma}_1(\hat{\boldsymbol{\theta}}), \quad \hat{\boldsymbol{\gamma}}_2 = \boldsymbol{\gamma}_2(\hat{\boldsymbol{\theta}}), \quad \hat{\boldsymbol{\gamma}}_1^* = \boldsymbol{\gamma}_1^*(\hat{\boldsymbol{\theta}}), \quad \hat{\boldsymbol{\gamma}}_2^* = \boldsymbol{\gamma}_2^*(\hat{\boldsymbol{\theta}}). \quad (\text{A.9})$$

To recover the scale-shifting intercept γ_0 and fixed effects $\{\sigma_i\}$, use $\sum_{i=1}^n \sigma_i = 0$:

$$\hat{\gamma}_0 = \frac{1}{nT} \sum_i \sum_t \left(|\hat{u}_{it}| - \hat{\boldsymbol{\theta}}' \mathbf{D}_t - \left[\hat{\boldsymbol{\gamma}}_1 + \hat{\boldsymbol{\theta}}' \mathbf{D}_t \cdot \hat{\boldsymbol{\gamma}}_1^* \right]' \mathbf{v}_{it} - \frac{1}{2} \left[\hat{\boldsymbol{\gamma}}_2 + \hat{\boldsymbol{\theta}}' \mathbf{D}_t \cdot \hat{\boldsymbol{\gamma}}_2^* \right]' \text{vec}(\mathbf{v}_{it} \mathbf{v}_{it}') \right), \quad (\text{A.10})$$

$$\hat{\sigma}_i = \frac{1}{T} \sum_t \left(|\hat{u}_{it}| - \hat{\gamma}_0 - \hat{\boldsymbol{\theta}}' \mathbf{D}_t - \left[\hat{\boldsymbol{\gamma}}_1 + \hat{\boldsymbol{\theta}}' \mathbf{D}_t \cdot \hat{\boldsymbol{\gamma}}_1^* \right]' \mathbf{v}_{it} - \frac{1}{2} \left[\hat{\boldsymbol{\gamma}}_2 + \hat{\boldsymbol{\theta}}' \mathbf{D}_t \cdot \hat{\boldsymbol{\gamma}}_2^* \right]' \text{vec}(\mathbf{v}_{it} \mathbf{v}_{it}') \right) \forall i. \quad (\text{A.11})$$

Step 3. For any given quantile index $0 < \tau < 1$ of interest, we next estimate the unconditional quantile of ε_{it} . From (3.2), we have the conditional quantile function of u_{it} :

$$\mathcal{Q}_u[\tau | \mathbf{v}_{it}, \mathbf{D}_t] = \left(\gamma_0 + \sum_{\kappa} \theta_{\kappa} D_{\kappa,t} + \left[\gamma_1 + \gamma_1^* \sum_{\kappa} \theta_{\kappa} D_{\kappa,t} \right]' \mathbf{v}_{it} + \frac{1}{2} \left[\gamma_2 + \gamma_2^* \sum_{\kappa} \theta_{\kappa} D_{\kappa,t} \right]' \text{vec}(\mathbf{v}_{it} \mathbf{v}_{it}') + \sigma_i \right) q_{\tau}. \quad (\text{A.12})$$

We therefore can estimate q_{τ} via a univariate quantile regression (with no intercept) via

$$\hat{q}_{\tau} = \arg \min_q \sum_i \sum_t \rho_{\tau} \left\{ \hat{u}_{it} - \left(\hat{\gamma}_0 + \hat{\boldsymbol{\theta}}' \mathbf{D}_t + \left[\hat{\boldsymbol{\gamma}}_1 + \hat{\boldsymbol{\theta}}' \mathbf{D}_t \cdot \hat{\boldsymbol{\gamma}}_1^* \right]' \mathbf{v}_{it} + \frac{1}{2} \left[\hat{\boldsymbol{\gamma}}_2 + \hat{\boldsymbol{\theta}}' \mathbf{D}_t \cdot \hat{\boldsymbol{\gamma}}_2^* \right]' \text{vec}(\mathbf{v}_{it} \mathbf{v}_{it}') + \hat{\sigma}_i \right) q \right\}, \quad (\text{A.13})$$

where $\rho_{\tau}\{\xi\} = \xi(\tau - \mathbb{I}\{\xi < 0\})$ is the check function, $\hat{u}_{it}$ is estimated in Step 1, and $[\hat{\boldsymbol{\theta}}', \hat{\gamma}_0, \hat{\boldsymbol{\gamma}}_1', \hat{\boldsymbol{\gamma}}_2', \hat{\boldsymbol{\gamma}}_1'^*, \hat{\boldsymbol{\gamma}}_2'^*]'$ and $\{\hat{\sigma}_i\}$ are estimated in Step 2.

With all unknown parameters now estimated, we can construct the estimator of the feasible analogue of the τ th conditional quantile of the log-cost in (3.5).

B Bias-Corrected Bootstrap Inference

To correct for finite-sample biases, we employ Efron's (1982) bias-corrected bootstrap percentile confidence intervals to conduct statistical inference. Bootstrap also significantly simplifies testing because, owing to a multi-step nature of our estimator, computation of the asymptotic variance of the parameter estimators is not trivial. Due to the panel structure of data, we use wild residual *block* bootstrap, thereby taking into account the potential dependence in residuals within each bank over time. The bootstrap algorithm is as follows.

- (i) Compute the estimator in Step 1. Save the estimated coefficients $[\hat{\eta}', \hat{\beta}_0, \hat{\beta}_1', \hat{\beta}_2', \hat{\beta}_1^{*'}, \hat{\beta}_2^{*'}]'$, location fixed effects $\{\hat{\lambda}_i\}$ and residuals $\{\hat{u}_{it}\}$.
- (ii) Generate bootstrap weights w_i^b for each cross-section/bank i from the two-point mass distribution:

$$w_i^b = \begin{cases} (1 + \sqrt{5})/2 & \text{with prob. } (\sqrt{5} - 1)/(2\sqrt{5}) \\ (1 - \sqrt{5})/2 & \text{with prob. } (\sqrt{5} + 1)/(2\sqrt{5}) \end{cases}. \quad (\text{B.1})$$

Next, for each observation (i, t) with $i = 1, \dots, n$ and $t = 1, \dots, T$, generate a new bootstrap disturbance as $u_{it}^b = w_i^b \times \hat{u}_{it}$.

- (iii) Construct a new bootstrap outcome variable: $c_{it}^b = \hat{\beta}_0 + \sum_{\kappa} \hat{\eta}_{\kappa} D_{\kappa,t} + [\hat{\beta}_1 + \hat{\beta}_1^* \sum_{\kappa} \hat{\eta}_{\kappa} D_{\kappa,t}]' \mathbf{v}_{it} + \frac{1}{2} [\hat{\beta}_2 + \hat{\beta}_2^* \sum_{\kappa} \hat{\eta}_{\kappa} D_{\kappa,t}]' \text{vec}(\mathbf{v}_{it} \mathbf{v}_{it}') + \hat{\lambda}_i + u_{it}^b$ for all $i = 1, \dots, n$ and $t = 1, \dots, T$.
- (iv) Recompute the Step 1 estimators in (A.3)–(A.5) using c_{it}^b in place of c_{it} to obtain bootstrap estimates of the location-function coefficients and fixed effects. Signify these by the superscript “ b .” Then, compute the bootstrap estimate of the residual $\hat{u}_{it}^b = c_{it} - \hat{\beta}_0^b - \sum_{\kappa} \hat{\eta}_{\kappa}^b D_{\kappa,t} - [\hat{\beta}_1^b + \hat{\beta}_1^{*b} \sum_{\kappa} \hat{\eta}_{\kappa}^b D_{\kappa,t}]' \mathbf{v}_{it} - \frac{1}{2} [\hat{\beta}_2^b + \hat{\beta}_2^{*b} \sum_{\kappa} \hat{\eta}_{\kappa}^b D_{\kappa,t}]' \text{vec}[\mathbf{v}_{it} \mathbf{v}_{it}'] - \hat{\lambda}_i^b$.
- (v) Reestimate the Step 2 estimator in (A.9) and the Step 3 estimator in (A.13) using $\hat{u}_{it}^b$ in place of $\hat{u}_{it}$ to obtain bootstrap estimates of the scale function coefficients and q_{τ} .
- (vi) Repeat bootstrap steps (ii)–(v) B times ($B = 500$ in this study). Use the empirical distribution of B bootstrap replicas of some estimand of interest (say, a coefficient or a quantile-specific function thereof such as cost subadditivity measure $\mathcal{S}_t^*$) to construct bias-corrected confidence intervals for this estimand.

To make matters concrete, let the (potentially, observation- and quantile-specific) estimand of interest be denoted by $\hat{\mathcal{E}}$. We can use the empirical distribution of $\{\hat{\mathcal{E}}^1, \dots, \hat{\mathcal{E}}^B\}$ to estimate the bias-corrected two-sided $(1 - \alpha) \times 100\%$ confidence bounds for $\mathcal{E}$ as an interval between the $[a_1 \times 100]$ th and $[(1 - a_2) \times 100]$ th percentiles of the bootstrap distribution, where $a_1 = \Phi(2\hat{z}_0 + \Phi^{-1}(\alpha/2))$ and $a_2 = \Phi(2\hat{z}_0 + \Phi^{-1}(1 - \alpha/2))$ with $\Phi(\cdot)$ being the standard normal cdf along with its quantile function $\Phi^{-1}(\cdot)$. Parameter $\hat{z}_0 = \Phi^{-1}(\#\{\hat{\mathcal{E}}^b < \hat{\mathcal{E}}\}/B)$ is a bias-correction factor measuring median bias, with $\#\{\mathcal{A}\}$ being a count function that returns the number of times event $\mathcal{A}$ is true. Naturally, to estimate the one-sided lower/upper $(1 - \alpha) \times 100\%$ confidence bound with bias correction, we respectively use the $[o_1 \times 100]$ th or $[(1 - o_2) \times 100]$ th percentiles of the bootstrap distribution, where $o_1 = \Phi(2\hat{z}_0 + \Phi^{-1}(\alpha))$ and $o_2 = \Phi(2\hat{z}_0 + \Phi^{-1}(1 - \alpha))$.