ESSAYS ON PRODUCTIVITY ANALYSIS

by

Jingfang Zhang

A dissertation submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Auburn, Alabama
December 11, 2021

Keywords: production function, productivity, cost function, scope Economies

Approved by

Emir Malikov, Chair, Affiliate Professor, Auburn University; Lee Professor of Economics, UNLV
Valentina Hartarska, Co-chair, Alumni Professor of Agricultural Economics
Ruiqing Miao, Associate Professor of Agricultural Economics
Joel Cuffey, Assistant Professor of Agricultural Economics

Abstract

This dissertation includes three essays conducting analyses of firm performance with the emphasis on estimating production technology, productivity, cost effectiveness and their relations with closely related aspects of firm behavior: exporting decisions, cross-firm learning and technology spillovers, local neighborhood influences, operational scope, etc. I contribute to the literature by extending and implementing methods for structural identification of firm-level production/cost functions and productivity to address issues of heterogeneity and endogeneity.

In the first chapter, I investigate the nexus between firm productivity and export behavior in a structural framework of firm production. My approach allows the firm's productivity to be affected not only by its own export behavior but also by that of its spatially proximate peers. The latter channel facilitated by cross-firm spillovers has been largely ignored in the literature. My model provides an internally consistent strategy for the measurement of productivity-boosting effects of exports that accommodates both channels. I apply it to Chilean manufacturing data from 1995-2007 and find significant evidence in favor of both the internal (within-firm) and external (cross-firm) effects of exporting.

The next chapter proposes a methodology to accommodate locational heterogeneity in production analysis. My approach is novel in that I explicitly model spatial variation in parameters in the production-function estimation. I accomplish this by allowing the parameters to be unknown functions of the firm's geographic location and estimate them via local kernel methods. This allows the production technology to vary across space, thereby accommodating neighborhood influences on firm production. Using this methodology, I study China's chemicals manufacturing in 2002-2004 and find that differences in technology are

the main source of the cross-location differential in total productivity in this industry.

The third chapter provides new and more robust evidence of scope economies in U.S. commercial banking. I improve upon the prior literature not only by analyzing the most recent data and accounting for banks' nontraditional non-interest-income-centered operations, but also in multiple methodological ways. I estimate a flexible time-varying-coefficient panel-data quantile regression model which accommodates three-way heterogeneity across banks. The results provide strong evidence in support of significantly positive scope economies across banks of virtually all sizes.

Acknowledgments

My four and half years at Auburn has been the best experience in my whole academic career. Many people have provided me their immense support, love, friendship and guidance and helped me to accomplish my Ph.D degree. I want to express my gratitude to them.

First and foremost, I would like to thank my parents and my sister for their immense support and love. I was born and raised in a small village in China surrounded by mountains, never dreamt of going to a big city, not to say abroad. My primary teacher once told my parents that I was stupid, but my parents never gave up on me. They have always supported me in a pursuit of my dream of an advanced study even though it has been a financial burden on them. My sister has always set a good example for me and helped me make the best decisions at critical moments. I would have never come to Auburn and gotten to meet such a great bunch of people without her constant encouragement.

Many professors have played an important role in the early years of my graduate studies. Professor Valentina Hartarska is my co-chair committee member, and I have worked with her on several projects. I am always impressed by her story-digging ability. I am encouraged and will always keep in mind what she once said, "Women can do better than men." Professor Ruiqing Miao is one of my committee members. He never hesitates to provide generous support to me. His rigorous research style, diligence, and love for research will continue to inspire me on my future academic path. I am also grateful to Professors Joel Cuffey and Wenying Li, who helped me set up mock interviews during my job search process. Wenying is a big picture thinker and taught me insightful perspectives on research and life. He always encourages me to reach out and socialize with people, which will be a lesson I will take for a lifetime.

A special thank you goes to my advisor, Professor Emir Malikov. Every time I saw and heard my Ph.D. fellows from various departments complaining about how irresponsible their advisors were and how much hardship they went through without guidance from advisors during their graduate studies, I could not help feeling that I was so lucky to have Dr. Malikov as my advisor. I doubt I would have successfully and smoothly finished my doctorate without his support, trust, and confidence in me. I am very grateful and delighted to have worked with Dr. Malikov during these years. He taught me how to do research, paved the way for me, and showed me the right direction for research. He sets an example of a good advisor, and I will pass on what I learned from him to my students in the future. I very much look forward to continuing collaboration with him for years to come.

Lastly, I want to express my appreciation to my partner, Yunei Hua, who makes my life colorful. His love, accompany, and encouragement are a source of my strength. We fell in love when we both had just come to Auburn. Auburn is a witness of our love story. I will miss Auburn and the people I met here!

Contents

Abstract 2

Acknowledgments 4

List of Tables 8

List of Figures 9

1 Learning by Exporting and from Exporters in Chilean Manufacturing **10**

1.1 Introduction . . . 10

1.2 Conceptual Framework . . . 15

1.3 Empirical Strategy . . . 20

1.4 Data . . . 25

1.5 Results . . . 26

1.6 Conclusion . . . 33

2 Accounting for Cross-Location Technological Heterogeneity in the Measurement of Operations Efficiency and Productivity **36**

2.1 Introduction . . . 36

2.2 Locational Heterogeneity in Production . . . 43

2.3 Methodology . . . 47

2.3.1 Proxy Variable Identification . . . 49

2.3.2 Semiparametric Estimation . . . 52

2.4 Locational Productivity Differential Decomposition . . . 55

2.5 Empirical Application . . . 57

2.5.1 Data . . . 57

2.5.2 Estimation Results . . . 59

2.6 Concluding Remarks . . . 70

3 Off-Balance-Sheet Activities and Scope Economies in U.S. Banking **72**

3.1 Introduction . . . 72

3.2 Theory of Multi-Product Costs . . . 77

3.3 Empirical Model . . . 80

3.3.1 Estimation Procedure 85
3.4 Data 88
3.5 Empirical Results 90
3.5.1 Scope Economies 91
3.5.2 Scale Economies 98
3.5.3 Technological Change 99
3.6 Conclusion 101

Appendices **112**

A Appendix to Chapter 2 112
A.1 The SAR Production-Function Models 112
A.2 Translog Technology 122
A.3 Finite-Sample Performance of the Estimator 125
A.4 Bootstrap Inference 127
A.5 Specification Test of Location Invariance 132
B Appendix to Chapter 3 136
B.1 Cost Elasticities 136
B.2 Bias-Corrected Bootstrap Inference 137

List of Tables

1.1 Data Summary Statistics, 1995–2007 26
1.2 Production Function Parameter Estimates 27
1.3 The LBE and LFE Productivity Effect Estimates 31
1.4 Estimates of the LBE and LFE Functions 32

2.1 Data Summary Statistics 58
2.2 Input Elasticity Estimates 61
2.3 Locationally Varying Returns to Scale Estimates 62
2.4 Productivity Process Coefficient Estimates 65
2.5 Locational Productivity Differential Decomposition 67

3.1 Data Summary Statistics 89
3.2 Cost Subadditivity Estimates 92
3.3 p-Values for the Stochastic Dominance Tests for Scope Economies Across Cost Quantiles 97
3.4 Returns to Scale Estimates 99
3.5 Technical Change 100
A.1 Second-Step Estimates of Locationally-Varying Parameters 126
A.2 Coverage Probability of the Two-Sided 95% Bootstrap Confidence Interval . . . 129
B.1 Cost Elasticity Estimates 136
B.2 Cost Elasticity Estimates (cont.) 137

List of Figures

1.1 Exporter Productivity Differential Estimates across Productivity Quantiles with the 95% Confidence Intervals 29
1.2 Distributions of log-Productivity by the Exporter Status 29

2.1 Spatial Distribution of Manufacturers of Chemicals in China, 2004–2006 38
2.2 Input Elasticity Estimates (Notes: Vertical lines correspond to location-invariant estimates) 61
2.3 Spatial Distribution of Returns to Scale Estimates (Notes: The color shade cutoffs correspond to the first, second (median) and third quartiles) 63
2.4 Productivity Process Coefficient Estimates (Notes: Vertical lines correspond to location-invariant estimates) 65
2.5 Locational Productivity Differential Decomposition Estimates Across Space (Notes: The color shade cutoffs correspond to the first, second (median) and third quartiles) . 69

3.1 The One-Sided 95% Lower Bounds (solid lines) of the Cost Subadditivity Point Estimates (scatter points) 94
3.2 Cost Subadditivity Estimates Across Cost Quantiles 96
3.3 Kernel Densities of Cost Subadditivity Estimates Across Cost Quantiles 96
A.1 Power of the Two-Sided 95% Bootstrap Confidence Interval (Vertical line corresponds to the true $\beta_K(s)$ value) 130

Chapter 1

Learning by Exporting and from Exporters in Chilean Manufacturing[*]

1.1 Introduction

Governments in both the developing and developed countries commonly pursue policies aimed at promoting exports. In addition to boosting aggregate demand, such policies are also routinely justified by arguing that domestic exporters benefit from export-driven productivity improvements via absorption of new technologies from abroad, learning of international best practices that lead to improved business processes, productivity enhancements driven by the exposure to more competition, scale effects, quality and variety effects, etc. The latter productiv-ity-enhancing mechanism is usually referred to as “learning by exporting” (LBE) (e.g., Clerides et al., 1998; Aw et al., 2000; Delgado et al., 2002; Baldwin and Gu, 2004; Van Biesebroeck, 2005; De Loecker, 2007; De Loecker, 2013). Such export-related productivity gains are facilitated by the firm’s *own* direct access to foreign customers, partners and rivals. However, not only can the domestic firms learn from their own export experiences, but they also can learn from their local *peers* who engage in exports, and this indirect learning opportunity is available to both exporters and non-exporters. These external export-driven productivity spillovers are a type of cross-firm peer effects, which we refer to as “learning from exporters” (LFE), and effectively capture secondary productivity effects

[*] This chapter is based on Zhang and Malikov (2019).

of export engagement. Such cross-firm spillovers may arise due to labor turnover, learning by imitation, customer-supplier discussions, etc. (see Greenaway et al., 2004; Sala and Yalcin, 2015). For instance, the movement of labor from the exporting firms to other domestic firms may facilitate the dispersion of tacit knowledge about more innovative/efficient foreign technologies or the institutional knowledge about foreign markets, which may help these firms improve their productivity as well.

Taking the indirect cross-firm productivity effect of export engagement for granted—as customarily done in the literature on the productivity–exports nexus—will likely underestimate the total productivity benefits of exporting. Besides, omitting this important mechanism may also contaminate the measurement of the more traditional LBE effect on firm productivity because it leads to an endogeneity-inducing omitted variable bias. In this paper, we extend De Loecker (2013) to develop a unified empirical framework for productivity measurement that explicitly accommodates both the direct LBE channel taking place *within* the firm as well as the indirect LFE channel working *between* firms, which we then apply to a panel of manufacturing plants in Chile in 1995–2007.

While the literature on the (within-firm) learning-by-exporting effects is rather well-established (e.g., Kunst and Marin, 1989; Aw and Hwang, 1995; Bernard and Jensen, 1999; Baldwin and Gu, 2003; Greenaway and Kneller, 2004; Keller, 2004; Blalock and Gertler, 2004; De Loecker, 2007; De Loecker, 2013; Wagner, 2007; Salomon and Jin, 2008; Park et al., 2010; Aw et al., 2011; Kasahara and Lapham, 2013; Manjón et al., 2013), the empirical analysis of external effects of exporting on *productivity* in the industry beyond the exporter firm is practically non-existent. The existing work on "export spillovers" focuses mainly on how the average export participation in the industry affects the export status or the marginal cost of non-exporters nearby (e.g., Aitken et al., 1997; Clerides et al., 1998; Bernard and Jensen, 2004; Greenaway et al., 2004; Greenaway and Kneller, 2008; Koenig, 2009; Koenig et al., 2010; Alvarez et al., 2013; Poncet and Mayneris, 2013). To our knowledge, the productivity implications of export spillovers in the domestic industry have not been studied empirically except for the two attempts (see below), both of which employ empirical strategies that are not only overly restrictive or "internally inconsistent" in their modeling of firm productivity (or both) but are also seriously hindered by well-known identification problems associated with the production function estimation.

Using the older Chilean manufacturing data, Alvarez and López (2008) also attempt to estimate productivity effects of export spillovers. They do so in two steps whereby they first estimate unobserved firm productivity via standard proxy methods while ignoring the dependence of the former on exports under the assumption of *exogenous* first-order Markov evolution of productivity and then examine spillover effects in a second step by (linearly) regressing the already estimated firm productivity on the average export spillover exposure. Taken at its face value, such a second-step analysis is problematic because it contradictorily postulates the existence of an endogenous exporting-productivity relationship that is at odds with the assumption about firm productivity being purely autoregressive in the first stage. Consequently, this approach cannot provide structurally meaningful interpretation of externality effects of exporting on productivity, also in part due to its inability to distinguish between different data-generating processes, including the one with an "external learning" mechanism, that all can give rise to a positive correlation between the export orientation of the industry and firm productivity (also see De Loecker, 2013).

We avoid this internal inconsistency in our model by explicitly accounting for potential export spillover-induced productivity improvements during the estimation of firm productivity, which is partly what enables us to consistently estimate the LFE effects along with other production-function components. Alternatively, the estimates of both the productivity and the export effects thereon would have suffered from the omitted variable bias due to omission of the relevant measurement of the peers' exporting activities from the productivity-proxy function that are correlated with the firm's own export behavior as well as quasi-fixed inputs (via its latent productivity). Thus, empirical findings of external export spillovers based on a two-step estimation procedure may be spurious. Furthermore, measuring the firm's spillover exposure using the average export intensity of all firms in the industry *including* a recipient of such externality effects, as pursued by Alvarez and López (2008), precludes the identification of purely external spillover effects of exporting on productivity by conflating it with the direct LBE effect of the firm's own export experience thereby, in the end, yielding a total of the two effects with no indication of their relative significance. In contrast, we distinguish between the firm's own export engagement and that of its peers, with the second measure capturing a purely external characteristic of the industry that the firm faces.

We also significantly improve upon Wei and Liu (2006). While employing a one-step analysis that recognizes the existence on both the LBE and LFE effects on productivity during the estimation, they do so by restrictively assuming that productivity effects of exporting are constant (as implied by linearity) which they estimate by including the export variables directly into the Cobb-Douglas production function. Besides linearity that rules out heterogeneous effects, such an approach is problematic because it assumes that the relationship between exporting and productivity is deterministic and, more importantly, it implies the possibility for a unit-elastic substitution between inputs and export variables (see De Loecker, 2013, for more on these points). Lastly, Wei and Liu (2006) proceed to estimate the production function via OLS with no account for the endogeneity problem associated with the correlation of input allocations (and potentially, the firm's export orientation) with the innovation in productivity (Griliches and Mairesse, 1998a).

In this paper, we contribute to the literature by robustly measuring and testing the direct learning-by-exporting and indirect learning-from-exporters effects on productivity in a consistent structural framework of firm production. Building upon Doraszelski and Jaumandreu (2013) and De Loecker (2013), we formalize the evolution of firm productivity as an endogenous export-controlled process, where we explicitly accommodate two potential channels —internal and external (direct and indirect)—by which exporting may impact future productivity of domestic firms. Specifically, we allow the firm's productivity be affected not only by its own export behavior but also by that of its spatially proximate peers in the industry. This allows a simultaneous, internally consistent identification of firm productivity and the corresponding LBE and LFE effects. Our identification strategy utilizes the structural link between the parametric production function and the firm's first-order condition for static inputs which helps us circumvent Ackerberg et al.'s (2015) and Gandhi et al.'s (2018) non-identification critiques of conventional proxy-based productivity estimators à la

Olley and Pakes (1996) and Levinsohn and Petrin (2003). In addition, owing to the nonparametric treatment of the firm productivity process, our model enables us to accommodate heterogeneity in productivity effects of exporting across firms. This also lets us explore potential nonlinearities in the LBE and LFE effects whereby they can interact with each other as well as, more importantly, with the firm's own productivity thus allowing for conditioning on the learning firm's absorptive capacity. To this end, not only do we provide a more comprehensive picture of the productivity effects of exporting, but we do so in a robust way by dealing with the internal inconsistency and non-identification problems prevalent in the earlier literature.

We study productivity-enhancing effects of exporting using plant-level data on Chilean manufacturers during the 1995–2007 period, with exporters accounting for 21% of the sample. Using our semiparametric methodology, we find that exporters enjoy a statistically significant productivity premium over non-exporters along the entire distribution of productivity. We find significant evidence in favor of both the LBE and LFE effects. Overall, the LBE productivity effect is statistically significant for 93% of all firms in our data set, although the results suggest that that the bulk of a productivity boost attributable to (internal) learning from exporting takes place immediately after the domestic firm engages in exports. On average, the size of the LFE effect is comparable to that of LBE. However, at the observation level, the LFE effect is significantly non-zero for 69% of plants only, thus suggesting that the indirect cross-firm productivity-boosting effect of exporting is less prevalent in manufacturing than the direct learning effect taking place within the firm. We also document that the LFE effect is stronger for the firms who also export themselves. This empirical evidence therefore suggests that exporters benefit from the exposure to peer exporters in their local industry more than do non-exporters, plausibly because there may be complementarities between internal/direct and external/indirect learning from export experiences. We also document that less productive plants benefit more from the export-driven learning (via both the internal or external channels) thereby suggesting that more productive plants may have less absorptive capacity to learn from their own export experiences as well as to absorb knowledge from their exporting peers. We also find that the more export-oriented firm is, the less (more) it learns by exporting (from exporters). A greater exposure to exporters helps plants absorb productivity improvements from their own export behavior while shows no significant effect on learning from exporters indicating that export spillovers in the industry improve plant productivity at a constant rate.

The rest of paper is organized as follows. Section 1.2 presents the conceptual framework. Section 1.3 describes our identification and estimation procedure. The data are discussed in Section 3.4. We report the empirical results in Section 3.5. Section 3.6 concludes.

1.2 Conceptual Framework

Consider the firm $i(=1,\ldots,n)$ at time $t(=1,\ldots,T)$. Following the convention in the productivity literature (e.g., Olley and Pakes, 1996; Blundell and Bond, 2000; Levinsohn and Petrin, 2003; De Loecker and Warzynski, 2012; Doraszelski and Jaumandreu, 2013; Ackerberg et al., 2015; Konings and Vanormelingen, 2015; Jin et al., 2019), we assume that the firm employs physical capital K_{it}, labor L_{it} and an intermediate input such as materials M_{it} to produce the output Y_{it} via the Cobb-Douglas production technology subject to the Hicks-neutral productivity:

$$Y_{it} = A_0 K_{it}^{\alpha_K} L_{it}^{\alpha_L} M_{it}^{\alpha_M} \exp\left\{\omega_{it} + \eta_{it}\right\}, \tag{1.2.1}$$

where A_0 is a scalar constant; $(\alpha_K, \alpha_L, \alpha_M)'$ are the input elasticities; ω_{it} is the firm's persistent productivity which is known to the firm at time t but unknown to an econometrician; and η_{it} is a random *i.i.d.* productivity shock such that $E[\eta_{it}|\mathcal{I}_{it}] = E[\eta_{it}] = 0$, where $\mathcal{I}_{it}$ is the ith firm's information set in period t.[1]

As in the productivity literature (e.g., Gandhi et al., 2020; Tsionas and Mallick, 2019; Hou et al., 2020), physical capital K_{it} and labor L_{it} are said to be subject to adjustment frictions (e.g., time-to-install, hiring and training costs), and the firm optimizes them dynamically at time $t-1$ rendering these inputs predetermined quasi-fixed state variables. Materials M_{it} is a freely varying input and is determined by the firm statically at time t. Both K_{it} and L_{it}

[1] We have also experimented with adding the time trend and its square term into the production function to control for temporal change. Our findings about the LBE and LFE productivity effects are largely unchanged and continue to hold. For simplicity sake, we have opted for a more parsimonious specification.

follow their respective laws of motion:

$$K_{it} = I_{it-1} + (1-\delta)\,K_{it-1} \quad \text{and} \quad L_{it} = H_{it} + L_{it-1}, \tag{1.2.2}$$

where I_{it}, H_{it} and δ respectively denote the gross investment, net hiring and the capital depreciation rate of the firm i in period t. We assume that the risk-neutral firm faces perfectly competitive output and input markets and seeks to maximize a discounted stream of the expected life-time profits subject to its state variables and expectations about the market structure variables including prices that are common to all firms.

In this paper, our principal interest is in the measurement of internal and external productivity effects of exporting in a domestic industry. Instead of modeling the firm's export behavior in a discrete fashion by focusing on its "status" as popularly done in the literature (e.g., Blalock and Gertler, 2004; Van Biesebroeck, 2005; Amiti and Konings, 2007; Kasahara and Lapham, 2013), we formalize the firm's exporting in a richer, continuous framework along the lines of De Loecker (2007) and Malikov et al. (2020). Specifically, we rely on the firm's export intensity as a measure of its own export behavior as well as to model its exposure to peer exporters in the industry. Let $X_{it} \in [0,1]$ denote the firm's export intensity defined as the nominal share of its total output produced for the export abroad, with its boundary values corresponding to wholly domestic and fully export-oriented firms. Building on Malikov et al. (2020), we conceptualize the firm's exporting decisions as the choice of the degree of its export orientation subject to delay due to costly adjustments. For instance, exporters may face irreversible adjustment costs such as time for and cost of finding new intermediaries/buyers abroad, contract (re)negotiations, obtaining new permits, etc. That is, we assume that the firm's decision to change the degree of its export orientation in period t (or leave it unchanged) is made at time $t-1$; ergo, the firm's export intensity X_{it} is predetermined.[2] Essentially, export adjustments are treated as an investment-like decision.

[2] Van Biesebroeck (2005) and De Loecker (2013) make similar assumptions about quasi-fixity of the export variable.

Hence, the firm's export intensity X_{it} evolves according to the following dynamic process:

$$X_{it} = \mathscr{X}_{it-1} + X_{it-1}, \tag{1.2.3}$$

where $\mathscr{X}_{it}$ is an endogenous adjustment in the degree of the firm's export orientation.

We next formalize the productivity effects of exports. We do so by extending De Loecker's (2013) framework to accommodate not only the more traditional direct LBE effects but also to allow for indirect effects via learning from the exporting peers. That is, we explicitly model two potential channels—internal and external (direct and indirect)—by which exporting may impact productivity of domestic firms.

The first channel, referred to as "learning by exporting," takes place *within* the firm internally and is commonly attributed to the exporter firm's absorption of new technologies from abroad, learning of international best practices that lead to improved manufacturing processes, productivity enhancements driven by the exposure to more competition, scale effects, quality and variety effects, etc. These export-related productivity gains are facilitated by the firm's *own* direct access to foreign customers and rivals.

The second channel is less obvious and oftentimes left unaccounted for in the literature. Domestic firms (both the exporters and non-exporters) can learn not only from their own export behavior but also indirectly from their exporting *peers*'. These external export-driven productivity spillovers are a type of *cross-firm* peer effects, which we refer to as "learning from exporters," and effectively capture secondary productivity effects of export engagement. Such cross-firm spillovers may arise due to, say, labor turnover, learning by imitation or customer-supplier discussions (see Greenaway et al., 2004; Sala and Yalcin, 2015). For instance, by monitoring successful exporting peers' market behavior both domestically and in the foreign markets, domestic firms can imitate and then adopt their business strategies to boost own productivity. Alternatively, the movement of labor from the exporting firms to other domestic firms may facilitate the dispersion of tacit knowledge about more innovative/efficient foreign technologies and better business practices or the institutional knowledge about foreign markets, which may help the hiring firms increase their productivity.

To capture export-driven productivity spillovers, we proxy each firm's exposure to exporters in the industry using the average export intensity of its spatially proximate *peers* operating in the same industry defined as

$$\overline{X}_{it} = \sum_{j \neq i} p_{ijt} X_{jt}, \tag{1.2.4}$$

where $\{p_{ijt};\ j(\neq i) = 1\ldots,n\}$ are the peer-firm weights identifying exporters in the firm i's industry and spatial locality in period t. More concretely, we construct the peer connection weights as $p_{ijt} = 1(j \in \mathscr{L}_{it}) \big/ \sum_{k(\neq i)=1}^{n} 1(k \in \mathscr{L}_{it})$, where $\mathscr{L}_{it}$ denotes a set of firms that are in the same industry and geographical region as is the firm i in time period t. This definition is based on the conventional argument that geographical proximity and industry play a central role in productivity spillovers. For example, it is technologically easier and less costly for firms to monitor and mimic strategies of other exporters that operate within the same industry and region. This is also in line with the literature on export spillovers (e.g., see Bernard and Jensen, 2004; Greenaway and Kneller, 2008; Koenig, 2009; Koenig et al., 2010; Poncet and Mayneris, 2013). We weigh all peers equally, given that the export intensity is already measured relative to the scale of production.

Note that our export exposure measure is firm-specific because it excludes the ith firm. Thus, $\overline{X}_{it}$ captures the *external* export orientation of the local industry which the firm i is exposed to. This measure varies across both the firms and time. While closely related, $\overline{X}_{it}$ is therefore not the "grand" industry average but the peer average in the industry. As discussed below, this distinction is crucial for the separable identification of the LBE and LFE effects.

We model firm productivity evolution as a controlled first-order Markov process, whereby we allow the firm i to improve its productivity not only via learning by exporting but also via learning from the exporting peers. Generalizing Doraszelski and Jaumandreu's (2013) and De Loecker's (2013) formulations to include cross-firm effects of exporting, we specify the following productivity process:

$$\omega_{it} = h\left(\omega_{it-1}, X_{it-1}, \overline{X}_{it-1}\right) + \zeta_{it}, \tag{1.2.5}$$

where $h(\cdot)$ is the conditional mean function of ω_{it}; and ζ_{it} is a random innovation in persistent productivity that is unanticipated by the firm at period $t-1$: $E\left[\zeta_{it}|\mathscr{I}_{it-1}\right] = E\left[\zeta_{it}|\omega_{it-1}, X_{it-1}, \overline{X}_{it-1}\right] = E\left[\zeta_{it}\right] = 0$.

The evolution process in (1.2.5) implicitly assumes that both the internal and external learning is a costly process which is why the dependence of ω_{it} on controls is lagged implying that the export-driven improvements in firm productivity take a period to materialize. Further, motivated by the literature on the productivity effects of exporting (e.g., Van Biesebroeck, 2005; De Loecker, 2013; Malikov et al., 2020), in $E[\zeta_{it}|\mathscr{I}_{it-1}] = 0$ we assume that, due to adjustment costs, the firm does not experience immediate changes in its export orientation in light of a productivity shock. This structural timing assumption about the arrival of ζ_{it}, which renders both X_{it-1} and $\overline{X}_{it-1}$ predetermined with respect to a random innovation at time t, helps identify both the direct learning and external spillover effects. The LBE and LFE effects can then be measured as $LBE_{it} = \partial E\left[\omega_{it}|\cdot\right] \big/ \partial X_{it-1}$ and $LFE_{it} = \partial E\left[\omega_{it}|\cdot\right] \big/ \partial \overline{X}_{it-1}$, respectively.

Since, in our productivity process (1.2.5), the exporting enters the conditional mean of productivity via two variables, of natural interest is the ability of our model to separate the direct LBE effect from the indirect LFE spillovers. Using simple calculus we can show that, owing to the definition of the average *peer* export orientation which excludes the export information pertaining to the ith firm:

$$\frac{\mathrm{d}h(\cdot)}{\mathrm{d}X_{it-1}} = \underbrace{\frac{\partial h(\cdot)}{\partial X_{it-1}}}_{LBE_{it}} + \underbrace{\frac{\partial h(\cdot)}{\partial \overline{X}_{it-1}}}_{LFE_{it}} \times \frac{\partial \overline{X}_{it-1}}{\partial X_{it-1}} = LBE_{it}, \tag{1.2.6}$$

because $\partial \overline{X}_{it-1} \big/ \partial X_{it-1} = \partial \sum_{j \neq i} p_{ijt} X_{jt-1} \big/ \partial X_{it-1} = 0$. Thus, LBE_{it} is *separably* identifiable. Intuitively, all observable variation in the expected productivity due to the change in the firm's own export intensity is attributable to the direct learning effect because its exporting does not immediately affect the export behavior of its peers. Obviously, the separability of the two effects would be impossible if, in place of the average *peer* export intensity, we would have used the total average of *all* firms as oftentimes done in the literature (e.g., Alvarez and López, 2008).

Lastly, owing to the unspecified nonparametric form of the conditional mean of ω_{it} in (1.2.5), we are able to obtain observation-specific estimates of the LBE and LFE effects thus allowing for potential cross-firm heterogeneity in the link between exporting and productivity. This also enables us to explore potential nonlinearities in the productivity effects of exporting whereby they can interact with each other as well as with the firm's own productivity.

1.3 Empirical Strategy

Estimating productivity using ordinary least squares regression would results in a simultaneity bias due to the dependence of inputs (regressors in the production function) on unobserved firm productivity ω_{it} because the latter is a part of the firm's information set $\mathscr{I}_{it}$ based upon which it makes optimal input allocation decisions. This omitted variable bias is also known as a "transmission bias" (Griliches and Mairesse, 1998a). A control-function-based method proposed by Olley and Pakes (1996) and extended by Levinsohn and Petrin (2003) tackles this endogenous problem by proxying for unobservable ω_{it} via the observable static intermediate input M_{it} and then using weakly exogenous higher-order lags of inputs to instrument for endogenous freely varying inputs. Recently, this methodology has been critiqued for the lack of identification due to perfect functional dependence between freely varying inputs and self-instrumenting quasi-fixed factors (Ackerberg et al., 2015) and violation of the "order condition" in the instrumentation of these endogenous freely varying inputs (Gandhi et al., 2020). As a solution, Gandhi et al. (2020) have suggested employing the information contained in the first-order condition for static inputs to identify both the production function and latent firm productivity. However, because their procedure is fully nonparametric, its implementation is three-stage and quite computationally burdensome, especially in its requirement to integrate the estimated static input elasticity function at each observation in order to recover the unknown production function. In this paper, we therefore rely on Malikov and Zhao's (2019) more easy-to-implement semiparametric adaptation of the Gandhi et al. (2020) methodology (which we modify to suit our research question) that utilizes the prespecified parametric form of the production function to derive the proxy function. This is similar to the idea pursued by Doraszelski and Jaumandreu (2013). The semiparametric adaptation can significantly ease the demand on data as well as the computational burden of estimation.

Identification.—Consider the firm's optimality condition for materials. Since the intermediate input M_{it} is freely varying and thus affects profits only in the current period, the firm's restricted expected profit maximization problem with respect to M_{it} is as follows:

$$\max_{M_{it}} P_t^Y A_0 K_{it}^{\alpha_K} L_{it}^{\alpha_L} M_{it}^{\alpha_M} \exp\{\omega_{it}\}\theta - P_t^M M_{it}, \tag{1.3.1}$$

where P_t^Y and P_t^M respectively denote the output and material input price, both of which are competitively determined. The constant θ is defined as $\theta \equiv E\left[\exp\{\eta_{it}\} \mid \mathscr{I}_{it}\right]$.

Taking the log-ratio of the first-order condition with respect to M_{it}

$$\alpha_M P_t^Y A_0 K_{it}^{\alpha_K} L_{it}^{\alpha_L} M_{it}^{\alpha_M - 1} \exp\{\omega_{it}\}\theta = P_t^M \tag{1.3.2}$$

and the production function in (1.2.1) gives

$$\ln\left(S_{it}^M\right) = \ln\left(\alpha_M \theta\right) - \eta_{it}, \tag{1.3.3}$$

where $S_{it}^M \equiv \frac{P_t^M M_{it}}{P_t^Y Y_{it}}$ is the intermediate input share of output. Thus, we can identify a composite constant $\alpha_M\theta$ from the unconditional moment $E\left[\eta_{it}\right] = 0$, from where we have that

$$\ln\left(\alpha_M \theta\right) = E\left[\ln\left(S_{it}^M\right)\right]. \tag{1.3.4}$$

We can also identify θ on its own via

$$\theta \equiv E\left[\exp\{\eta_{it}\}\right] = E\left[\exp\left\{\ln\left(\alpha_M\theta\right) - \ln\left(S_{it}^M\right)\right\}\right] = E\left[\exp\left\{E\left[\ln\left(S_{it}^M\right)\right] - \ln\left(S_{it}^M\right)\right\}\right], \tag{1.3.5}$$

with equation (1.3.4) used to substitute for $\ln\left(\alpha_M\theta\right)$ in the third equality.

Combining (1.3.4) and (2.3.8), we identify the firm's material elasticity α_M as

$$\alpha_M = \exp\left\{E\left[\ln\left(S^M_{it}\right)\right]\right\} / E\left[\exp\left\{E\left[\ln\left(S^M_{it}\right)\right] - \ln\left(S^M_{it}\right)\right\}\right], \tag{1.3.6}$$

where it is a unique function of the first moments of data.

To identify the rest of production function as well as latent firm productivity, we take the log of (1.2.1) on both sides to obtain

$$y_{it} = \alpha_0 + \alpha_K k_{it} + \alpha_L l_{it} + \alpha_M m_{it} + \omega_{it} + \eta_{it}, \tag{1.3.7}$$

where $\alpha_0 \equiv \ln A_0$; and the lower-case variables correspond to the log form of the respective upper-case variables. Exploiting the Markov process of ω_{it} in (1.2.5) and bringing the already identified material elasticity α_M to the left-hand side, we rewrite (1.3.7) as follows:

$$y^*_{it} = \alpha_K k_{it} + \alpha_L l_{it} + g\left(\omega_{it-1}, X_{it-1}, \overline{X}_{it-1}\right) + \zeta_{it} + \eta_{it}, \tag{1.3.8}$$

where $y^*_{it} = y_{it} - \alpha_M m_{it}$ is fully identified and can be treated as an observable, and $g(\cdot) \equiv h(\cdot) + \alpha_0$ is of unknown functional form.

Next, from equation (1.3.2) we derive the explicit form of the conditional demand function for M_{it}, which we then invert to proxy for the unobservable scalar ω_{it} in (1.3.8) in the spirit of material-based proxy estimators:

$$y^*_{it} = \alpha_K k_{it} + \alpha_L l_{it} + g\left(\left[m^*_{it-1} - \alpha_K k_{it-1} - \alpha_L l_{it-1}\right], X_{it-1}, \overline{X}_{it-1}\right) + \zeta_{it} + \eta_{it}, \tag{1.3.9}$$

where $m^*_{it-1} = \ln\left(\frac{P^M_{t-1}}{P^Y_{t-1}}\right) - \ln(\alpha_M \theta) - (\alpha_M - 1)\, m_{it-1}$ is also fully identified and treated as an observable. Since all regressors appearing in (1.3.9) including k_{it}, l_{it}, k_{it-1}, l_{it-1}, $m^*_{it-1}(m_{it-1})$, X_{it-1} and $\overline{X}_{it-1}$ are predetermined based on our structural assumptions, there is no endogenous covariate on the right-hand side of the equation. That is,

$$E\left[\zeta_{it} + \eta_{it} \middle| k_{it}, l_{it}, k_{it-1}, l_{it-1}, m^*_{it-1}(m_{it-1}), X_{it-1}, \overline{X}_{it-1}\right] = 0, \tag{1.3.10}$$

and the equation (1.3.9) is identified.

One remark is in order here. From (1.3.9)–(1.3.10), it is obvious that, if there were indeed non-zero export spillovers and we had failed to account for them in the firm's productivity evolution process, then $\overline{X}_{it-1}$ would have been omitted from the proxy function $g(\cdot)$ in (1.3.9) and, consequently, been absorbed, along with its interactions with other arguments of the proxy, into the error term. In the latter case, the error term would then contain variation from the firm's quasi-fixed inputs, its own export intensity as well as the average export orientation of its peers. Generally, these all would be correlated with quasi-fixed inputs and the export variable included as regressors thus violating the exogeneity condition. The model would be unidentified due to the omitted variable bias. This highlights the importance of embedding the external spillover channel into the analytical framework explicitly.

Lastly, we can recover latent firm productivity up to a constant:

$$\omega_{it} + \alpha_0 = y_{it} - \alpha_K k_{it} - \alpha_L l_{it} - \alpha_M m_{it} - \eta_{it}, \tag{1.3.11}$$

using the identified production-function parameters and the productivity shock.

Estimation Procedure.—The estimation is simple and involves a two-stage procedure. In the first stage, we estimate α_M via a sample counterpart of (1.3.6) constructed using sample averages computed from the raw data on material share:

$$\widehat{\alpha}_M = \exp\left\{\frac{1}{nT}\sum_i\sum_t \ln\left(S^M_{it}\right)\right\} \Big/ \left[\frac{1}{nT}\sum_i\sum_t \exp\left\{\left[\frac{1}{nT}\sum_i\sum_t \ln\left(S^M_{it}\right)\right] - \ln\left(S^M_{it}\right)\right\}\right]. \tag{1.3.12}$$

As a by-product, we also have $\ln\widehat{(\alpha_M\theta)} = \frac{1}{nT}\sum_i\sum_t \ln\left(S^M_{it}\right)$ and $\widehat{\eta}_{it} = \ln\widehat{(\alpha_M\theta)} - \ln\left(S^M_{it}\right)$. With these estimates in hand, we then construct estimates of y^*_{it} and m^*_{it} as $\widehat{y}^*_{it} = y_{it} - \widehat{\alpha}_M m_{it}$ and $\widehat{m}^*_{it-1} = \ln\left(\frac{P^M_{t-1}}{P^Y_{t-1}}\right) - \ln\widehat{(\alpha_M\theta)} - \left(\widehat{\alpha}_M - 1\right) m_{it-1}$, respectively.

The second-stage estimation requires the choice of an approximator for the unknown $g(\cdot)$. We use the popular second-order polynomial sieves (e.g., Gandhi et al., 2020).[3] Specif-

[3] We have also experimented with third-order polynomials, and the results are qualitatively similar except noisier, as expected.

ically, we approximate $g(\cdot)$ as follows:

$$\begin{aligned} g(\cdot) &\approx \Big(1, W_{it-1}(\boldsymbol{\alpha}), W^2_{it-1}(\boldsymbol{\alpha}), X_{it-1}, X^2_{it-1}, \overline{X}_{it-1}, \overline{X}^2_{it-1}, W_{it-1}(\boldsymbol{\alpha})X_{it-1}, W_{it-1}(\boldsymbol{\alpha})\overline{X}_{it-1}, X_{it-1}\overline{X}_{it-1}\Big)\boldsymbol{\gamma} \\ &= \boldsymbol{\lambda}_{it}(\boldsymbol{\alpha})'\boldsymbol{\gamma} \end{aligned}$$

where we let $\boldsymbol{\alpha} = (\alpha_K, \alpha_L)'$, $W_{it-1}(\boldsymbol{\alpha}) = \widehat{m}^*_{it-1} - \alpha_K k_{it-1} - \alpha_L l_{it-1}$, and $\boldsymbol{\gamma}$ is the unknown parameter vector.

We estimate (1.3.9) using a nonlinear least squares method to obtain the second-stage estimates of $(\alpha_K, \alpha_L)'$ and $\boldsymbol{\gamma}$:

$$\min_{\alpha_K, \alpha_L, \boldsymbol{\gamma}} \sum_i \sum_t \left(\widehat{y}^*_{it} - \alpha_K k_{it} - \alpha_L l_{it} - \boldsymbol{\lambda}_{it}(\alpha_K, \alpha_L)'\boldsymbol{\gamma}\right)^2. \tag{1.3.13}$$

With the estimated production-function parameters in hand, we can compute the productivity effects via $\widehat{LBE}_{it} = \partial\widehat{g}(\cdot)/\partial X_{it-1}$ and $\widehat{LFE}_{it} = \partial\widehat{g}(\cdot)/\partial\overline{X}_{it-1}$, where $\widehat{g}(\cdot) = \boldsymbol{\lambda}_{it}(\widehat{\boldsymbol{\alpha}})'\widehat{\boldsymbol{\gamma}}$.[4] We also recover ω_{it} up to a constant via $\widehat{\omega_{it} + \alpha_0} = y_{it} - \widehat{\alpha}_K k_{it} - \widehat{\alpha}_L l_{it} - \widehat{\alpha}_M m_{it} - \widehat{\eta}_{it}$.

Bootstrap.—For statistical inference, we employ accelerated biased-corrected percentile bootstrap confidence intervals proposed by Efron (1987), which can correct for finite-sample biases and control for higher moments (skewness) of the sampling distribution. Due to the panel structure of data, we employ wild residual block bootstrap, which can preserve the within-firm correlation in the data, to approximate the sampling distribution of the estimator. In addition, we bootstrap both stages jointly because the estimation in the second stage is based on the first-stage estimator.

Having obtained the bootstrap estimates of all parameters $\left\{\left(\widehat{\alpha}^b_K, \widehat{\alpha}^b_L, \widehat{\alpha}^b_M, \widehat{\boldsymbol{\gamma}}^b\right)';\ b = 1, 2, ...B\right\}$,[5] we then use them to construct the bootstrap replications of our main estimands of interest: $\left\{\widehat{LBE}^b_{it}\right\}$ and $\left\{\widehat{LFE}^b_{it}\right\}$, which we then use to construct accelerated biased-corrected percentile confidence intervals for each observation-specific LBE_{it} and LFE_{it}. For simplicity, let $\widehat{Z}$ represent an estimate of some functional of interest Z and $\{\widehat{Z}^b\}$ be the set of its bootstrap estimates. Then, a two-sided $(1-a)100\%$ confidence intervals for Z is the $[\alpha_1 \times 100]$th

[4] Note that the definitions of the LBE and LFE effects are based on the gradients of $h(\cdot)$ but, since $g(\cdot)$ and $h(\cdot)$ differ only by an additive constant, their gradients are equal.

[5] We set $B = 500$.

and $[\alpha_2 \times 100]$th percentiles of the empirical distribution of $\{\widehat{Z}^b\}$, where

$$\alpha_1 = \Phi\left(\widehat{q}_0 + \frac{\widehat{q}_0 + q_{\alpha/2}}{1 - \widehat{c}\left(\widehat{q}_0 + q_{\alpha/2}\right)}\right),\ \alpha_2 = \Phi\left(\widehat{q}_0 + \frac{\widehat{q}_0 + q_{(1-\alpha/2)}}{1 - \widehat{c}\left(\widehat{q}_0 + q_{(1-\alpha/2)}\right)}\right),$$

and Φ denotes the standard normal cdf; $q_{\alpha/2} = \Phi^{-1}(\alpha/2)$; $\widehat{q}_0$ denotes a bias-correction factor defined as $\widehat{q}_0 = \Phi^{-1}\left(\frac{\#\{\widehat{Z}^b < \widehat{Z}\}}{B}\right)$; $\widehat{c}$ denotes an acceleration parameter which, following the literature (e.g., Shao and Tu, 1995), we estimate via jackknife: $\widehat{c} = \sum_{j=1}^{J}\left(\sum_{s=1}^{J} \widehat{Z}^s - \widehat{Z}^j\right) / 6\left[\sum_{j=1}^{J}\left(\sum_{s=1}^{J} \widehat{Z}^s - \widehat{Z}^j\right)^2\right]^{3/2}$, where $\{\widehat{Z}^j\}$ are the jackknife estimates of Z.[6]

1.4 Data

Our data come from the Encuesta Nacional Industrial Anual (ENIA), a national industrial survey, conducted by the Chilean National Institute of Statistics annually. The sample period runs from 1995 to 2007. Manufacturing plants are classified into 22 industry groups according to the 2-digit International Standard Industry Classification (ISIC) code. The dataset contains information on plants from 13 regions including Tarapacá, Antofagasta, Atacama, Coquimbo, Valparaíso, Libertador Gral. Bernardo O'Higgins, Maule, Biobío, La Araucanía, Los Lagos, Aisén del Gral. Carlos Ibáñez del Campo, Magallanes and Chilean Antarctica, and the Santiago Metropolitan region. Though each observation represents a plant rather than a firm, single-plant establishments account for over 90% of total units (also see Pavcnik, 2002).

The total output is defined as the total revenue from the sale of products and work done. Capital is the fixed assets balance for buildings, machinery and vehicles at the end of a period. Materials are defined as the total expenditure on intermediate inputs consisting of raw materials and other intermediates. These three variables are measured in hundred thousands of pesos, and we deflate them using price deflators at the 4-digit ISIC level. We measure labor using the total number of people working at the plant. We drop observations that contain missing or negative values for these variables and exclude extreme outliers lying outside the interval between the 1th and 99th percentiles of these four variables. In the end,

[6] To account for the panel structure of data and to manage computational time, we use a delete-20T jackknife, i.e., we leave 20 cross-sections out to obtain jackknife estimates.

Table 1.1. Data Summary Statistics, 1995–2007

Variable	Mean	5th Perc.	Median	95th Perc.
Output (Y)	276.46	8.37	54.75	1348.35
Capital (K)	102.24	0.48	12.15	513.97
Labor (L)	49.74	6.00	22.00	196.00
Materials (M)	125.94	2.85	24.09	626.56
Export Intensity (X)	0.056	0.000	0.000	0.499
Exposure to Exporter ($\overline{X}$)	0.060	0.000	0.036	0.217
Exporter Status	0.206			

Notes: Y, K, M are measured in hundred thousands of real pesos. L is measured in the number of people. X is a unit-free proportion of firm's exports in total output. $\overline{X}$ is also a unit-free proportion.

our sample consists of 8,353 manufacturing plants with the total of 47,622 observations.

Export intensity is calculated as the nominal proportion of firm's exports in its total sales, ranging from 0 to 1 by construction. Out of all plants, exporters are 21%. As discussed earlier, we measure each plant's exposure to exporters using the average export intensity of its peers (excluding the plant in question). Peers are identified as operating in the same of 13 regions and the same 2-digit industry in the same period. The exposure variable also ranges between 0 and 1 by construction. Table 1.1 provides the summary statistics for our data.

1.5 Results

Our primary interest is in the estimates of productivity effects of exporting. Owing to a nonparametric form of the conditional mean of ω_{it}, we obtain observation-specific estimates of the LBE and LFE effects. Since these effects are defined as the gradients of the firm log-productivity with respect to export intensity or the average thereof both of which are proportions, the reported LBE_{it} (LFE_{it}) is a semi-elasticity measuring percentage changes in productivity per unit percentage point change in the firm's export intensity (the average external export orientation of the local industry that the firm faces). Lastly, both effects measure "short-run" productivity improvements per annum, which however can accumulate over the years owing to a persistent autoregressive nature of the firm's productivity evolution.

Production Function.—Before proceeding to the main discussion of productivity and the

Table 1.2. Production Function Parameter Estimates

Parameter	Point Estimate	Lower Bound	Upper Bound
Capital Elasticity	0.214	0.204	0.226
Labor Elasticity	0.456	0.435	0.476
Material Elasticity	0.288	0.284	0.291
Scale Elasticity	0.957	0.938	0.975

Notes: Reported are the input elasticity estimates along with their two-sided 95% lower and upper confidence bounds. Scale elasticity is the sum of capital, labor and material elasticities.

implications of exporting on the former, we first consider estimates of the production function parameters. Table 1.2 reports point estimates of the capital, labor and material elasticities along with the lower and upper bounds corresponding to two-sided 95% bootstrap percentile confidence intervals. Elasticities of capital, labor and material are all statistically significant at 0.29, 0.21 and 0.46, respectively. Scale elasticity defined as the sum of all three input elasticities is statistically significant at 0.96, indicating the decreasing return to scale, consistent with the profit-maximizing behavior.

Exporter Productivity Differential.—We begin by examining an overall cross-firm productivity differential across exporters and non-exporters in the Chilean manufacturing sector. The mean estimate of the (log)-productivity differential between exporters and non-exporters is 0.280 with the corresponding two-sided 95% bootstrap percentile confidence interval of (0.255, 0.308), which indicates that, on average, exporters are more productive than non-exporters. This is consistent with findings in the related literature on productivity of Chilean exporters (e.g., Alvarez and Lopez, 2005).

We further investigate if the plant's exporter status commands a productivity premium of a varying magnitude and significance at different points in the productivity distribution. We do so by estimating the following simple quantile regression:

$$\mathbb{Q}_{\tau}\left[\omega_{it}|\cdot\right] = \beta_{0,\tau} + \beta_{1,\tau}\text{EXP}_{it} \qquad \text{for } \tau \in (0,1),$$

where EXP_{it} is the exporter status indicator such that $1(X_{it} > 0)$. Employing the conditional quantile regression enables us to explore potential distribution heterogeneity in the exporter productivity differential. We estimate this model for the quantile index τ taking values from 0.2 to 0.8 (with the 0.05 increments) thereby focusing on the central portion of the productivity distribution.

Figure 1.1(a) plots the quantile regression estimates of the $\beta_{1,\tau}$ coefficient (the exporter productivity differential) against τ, along with the 95% confidence intervals. The solid horizontal line corresponds to the productivity differential estimated at the conditional mean. The quantile productivity differentials are all significantly positive and increasing with the quantile of ω, indicating that the productivity divergence between exporters and non-exporters is more prominent magnitude-wise among the more productive firms. Including controls for the plant size (proxied by the number of employees) as well as the region and year effects produces the same findings as can be seen in Figure 1.1(b).

For a more holistic look at the productivity differential between exporters and non-exporters, we also plot the kernel density of the firm log-productivity for exporters and non-exporters, as shown in Figure 1.2. This allows us to compare distributions of productivity estimates as opposed to merely focusing on marginal moments or quantiles. The figure indicates that exporters appear to enjoy a productivity premium over non-exporters distribution-wise. To support this visual evidence, we do a formal test to check if exporters are more productive than non-exporters along the entire distribution of productivity. We utilize a generalization of the Kolmogorov-Smirnov test proposed by Linton et al. (2005) to test the (first-order) stochastic dominance of exporters' productivity over non-exporters'. This test permits variables to be estimated latent quantities as opposed to observables from the data and to also share dependence (in our case, the dependence due to their construction using the same set of parameter estimates). Specifically, let $G_1(\omega)$ and $G_0(\omega)$ denote the cumulative distribution functions of productivity $\omega \in \Omega$ for exporters and non-exporters, respectively. We then construct the null hypothesis that non-exporters' productivity is stochastically dominated by that of exporters as follows:

$$H_0: \sup_{\omega\in\Omega}\left[G_1(\omega) - G_0(\omega)\right] \leq 0 \quad \text{vs.} \quad H_1: \sup_{\omega\in\Omega}\left[G_1(\omega) - G_0(\omega)\right] > 0,$$

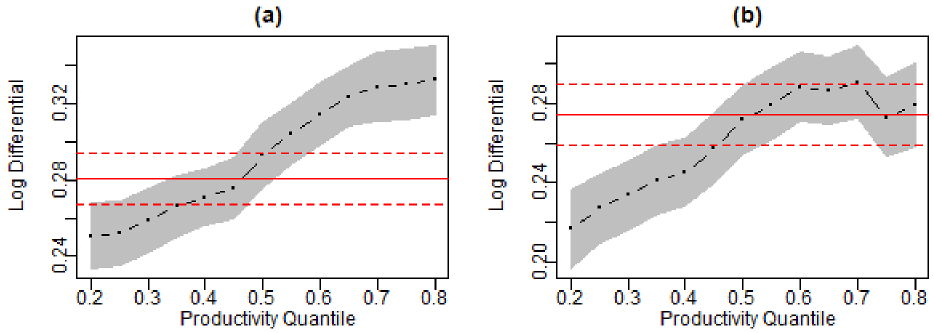

Figure 1.1. Exporter Productivity Differential Estimates across Productivity Quantiles with the 95% Confidence Intervals

Notes: Solid horizontal lines correspond to the productivity differentials estimated at the conditional mean.

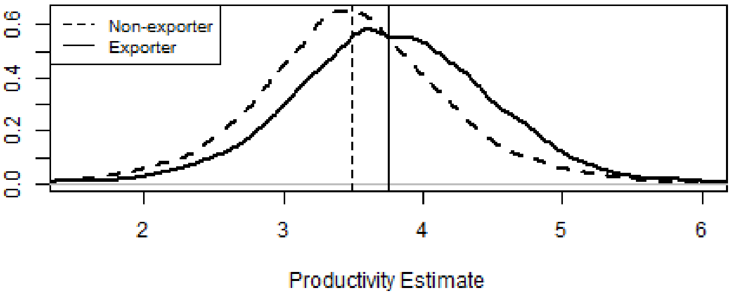

Figure 1.2. Distributions of log-Productivity by the Exporter Status

Notes: Vertical lines correspond to the respective sample means.

with the corresponding test statistic defined as

$$\mathscr{D} = \max_{1 \le j \le (n_1 + n_0)} \sqrt{\frac{n_1 n_0}{n_1 + n_0}} \left[G_{1,n_1} \left(\widehat{\omega}_j \right) - G_{0,n_0} \left(\widehat{\omega}_j \right) \right],$$

where G_{s,n_s} is the empirical distribution function of the estimated (log) productivity ω for the sth category of plants of the sample size n_s, with $s = \{0, 1\}$. Employing the sub-sampling procedure from Linton et al. (2005), we obtain the p-value for the test statistic of 0.7789.[7] Thus, we fail to reject the null hypothesis whereby non-exporters' productivity is stochastically dominated by that of exporters. Combined with the above discussion, we can conclude that exporters enjoy a statistically significant productivity premium over non-exporters along the entire distribution of productivity.

Learning by Exporting and from Exporters.—Next, we consider the within-firm evidence of productivity-enhancing effects of exporting. Table 1.3 reports a summary of point estimates of the LBE and LFE effects for all firms as well as for exporters and non-exporters only. We also test for the statistical significance of these effects at *each* observation. The shares of observations for which each of the two productivity effects of exporting is significant at the 5% level are provided in the last column of the table.

The LBE effect is estimated to average at 0.36 for the entire sample, indicating that a 1 percentage point increase in the firm's own export intensity raises its future productivity by 0.36%. The point estimates are between 0.32 and 0.45 within the inter-quartile range. Overall, the LBE productivity effect is statistically significant for 93% of firms in our data set. However, we do document notable differences in the magnitude and prevalence of LBE across exporters and non-exporters. For actual exporters ($X_{it} > 0$), the mean LBE effect is significant at 0.158, albeit the observation-specific point estimates are statistically non-zero only for 68% of the exporting firms. This contrasts starkly with the results for non-exporters. While the latter category of firms does not actually export, we still can evaluate the LBE effect on their future productivity at $X_{it} = 0$. Essentially, these estimates of the LBE effect

[7] We use r_n equidistant sub-sample sizes $B_n = \{b_1, \cdots, b_r\}$, where $b_1 = \left[\log\log n\right]$, $b_{r_n} = \left[n / \log\log n\right]$, and the number of unique sub-sample sizes is $r = 199$. For each b, we get a p-value. The reported is the mean of these p-values.

Table 1.3. The LBE and LFE Productivity Effect Estimates

	Point Estimates				Stat. Signif.
	Mean	1st Qu.	Median	3rd Qu.	(% Obs.)
			—Learning by Exporting—		
All	0.363	0.323	0.39	0.451	92.7
	(0.189, 0.555)	(0.139, 0.531)	(0.207, 0.603)	(0.25, 0.675)	
Exporters	0.158	–0.000	0.257	0.353	68.3
	(0.068, 0.254)	(–0.072, 0.067)	(0.117, 0.419)	(0.184, 0.546)	
Non-exporters	0.418	0.356	0.408	0.465	99.2
	(0.234, 0.643)	(0.170, 0.572)	(0.220, 0.628)	(0.259, 0.695)	
			—Learning from Exporters—		
All	0.324	0.141	0.296	0.457	68.5
	(0.116, 0.545)	(–0.059, 0.363)	(0.069, 0.52)	(0.166, 0.729)	
Exporters	0.508	0.182	0.399	0.778	73.9
	(0.302, 0.760)	(–0.013, 0.432)	(0.189, 0.618)	(0.508, 1.148)	
Non-exporters	0.275	0.132	0.28	0.42	67.1
	(0.027, 0.489)	(–0.088, 0.340)	(0.027, 0.497)	(0.124, 0.701)	

Notes: Reported is a summary of point estimates of the LBE and LFE effects tabulated by the firm's exporter status, with two-sided 95% bootstrap percentile confidence intervals in parentheses. The far right column reports the share of sample for which the observation-specific estimates are statistically significant at the 5% level.

are "counterfactual" and provide a measurement of how much non-exporters' productivity would have changed if they started exporting (marginally increased their export intensity from zero to a positive value). The average LBE estimate for non-exporters is 0.418 and significant. In fact, the point estimates of the LBE effect are statistically significant for virtually all non-exporters (99%). Magnitude-wise, the average effect size for non-exporters is about 2.6 times larger than that for active exporters. This is reasonable as the decision to export may go together with other firm-level actions that can enhance productivity, such as technology adoption, quality upgrading or R&D spending (e.g., Verhoogen, 2008; Aw et al., 2011; Bustos, 2011). Due to the lack of rich data, our LBE estimates are more of the "reduced-form" estimates that bundle different channels together. These findings reasonably suggest that the bulk of a productivity boost attributable to (internal) learning from exporting and probably other exporting related investment takes place immediately after the domestic firm engages in exports and thus gains access to new technology and business practices.

On average, the size of the LFE effect is comparable to that of LBE: the pooled mean estimate of the LFE effect is statistically significant at 0.32 (vs. 0.36). However, at the observation level, the LFE effect is significantly non-zero only for 69% of all firms, thus sug-

Table 1.4. Estimates of the LBE and LFE Functions

	LBE	LFE
ω_{it-1}	–0.1127	–0.3455
	(–0.212, –0.039)	(–0.575, –0.1338)
X_{it-1}	–0.9698	1.2104
	(–1.4543, –0.4346)	(0.5832, 2.0077)
$\overline{X}_{it-1}$	1.2104	0.2198
	(0.5832, 2.0077)	(–0.9592, 1.3766)

Notes: Reported are the parameter estimates for the LBE and LFE functions derived from the polynomial approximation of the conditional mean of ω_{it}, along with the two-sided 95% bootstrap percentile confidence intervals in parentheses.

gesting that the indirect cross-firm productivity-boosting effect of exporting is less prevalent in manufacturing than the direct learning effect taking place within the firm. We also interestingly document that the LFE effect is stronger for the firms who also export themselves. Namely, the average estimate of the LFE effect for the exporter firms is estimated at 0.508, whereas the corresponding estimate for non-exporters is half that at 0.275. In addition, non-zero learning from exporters is also more prevalent for exporters (74%) than it is for non-exporters (67%). The empirical evidence therefore suggests that exporters benefit from the exposure to peer exporters in their local industry more than do non-exporters. Plausibly, there may be complementarities between internal/direct and external/indirect learning from export experiences. This is reasonable because challenges associated with engagements in the foreign market can make exporters more motivated and pressured than their fully domestically-oriented non-exporting peers to improve further and more intensely so. Overall however, the external LFE effect is significant for most manufacturing plants.

Next, we explore heterogeneity in the productivity effects of exporting. Firms are highly heterogeneous across many dimensions including their productivity, the degree of their export orientation as well as the intensity of their exposure to other exporters in the industry. In what follows, we investigate if these characteristics influence the effect size of internal and external learning from exporting.

Recall that we obtain the estimate of the productivity effects of exporting via $\widehat{LBE}_{it} = \partial\widehat{g}(\cdot)/\partial X_{it-1}$ and $\widehat{LFE}_{it} = \partial\widehat{g}(\cdot)/\partial\overline{X}_{it-1}$, where we estimate $g\left(\omega_{it-1}, X_{it-1}, \overline{X}_{it-1}\right)$ using the second-order polynomial sieve approximation. Thus, by derivation, both $\widehat{LBE}_{it}$ and $\widehat{LFE}_{it}$ are the estimated linear functions of the "determinants" of firm productivity $\left(\omega_{it-1}, X_{it-1}, \overline{X}_{it-1}\right)'$. Table 1.4 reports the estimates of parameters on these three variables for LBE and LFE.

The coefficient estimates on ω_{it-1} for both LBE and LFE are significantly negative, indicating that the magnitude of these effects declines as firms get more productive. Thus, less productive plants benefit more from export-driven learning, be it an internal or external channel. This finding is largely in line with economic intuition whereby more productive plants would have less absorptive capacity to learn from their own export experiences as well as to absorb knowledge from their surroundings including the exporting peers. We also find that the firm's own export intensity X_{it-1} has a significantly negative effect on learning by exporting but a significantly positive effect on learning from exporters, indicating that less export-oriented plants improve more via learning by exporting and less via learning from exporters (and vice versa). Basically, this is indicative of the diminishing productivity return to the own export experience: an increase in the degree of firm's export orientation enhances its productivity at a decreasing rate. But we do not find such a pattern for the LFE effect. On the contrary, the more export-oriented the plant is, the higher the cross-firm export-driven productivity spillovers are, which buttresses our earlier discussion of potential complementarities between exporting and external learning from Table 1.3. Lastly, the results in Table 1.4 suggest that the average of peer export orientation in the local industry $\overline{X}_{it-1}$ has a significantly positive influence on learning by exporting, indicating that a greater exposure to exporters helps plants absorb productivity improvements from their own export behavior. At the same time, we find no significant effect of $\overline{X}_{it-1}$ on the LFE effect size, indicating that the average export participation in the industry improves future plant productivity at a constant rate.

1.6 Conclusion

Governments in both the developing and developed countries commonly pursue policies aimed at promoting exports. In addition to boosting aggregate demand, such policies are also routinely justified by arguing that domestic exporters benefit from export-driven productivity improvements. When studying these productivity effects, the existing literature mostly focuses on whether firms improve their performance by engaging in exports *themselves*, a mechanism called "learning by exporting," while largely neglecting a secondary channel whereby domestic firms can also learn from their exporting *peers* via cross-firm spillovers. This indirect learning opportunity, which we refer to as "learning from exporters," is available to both exporters and non-exporters. Omitting this important mechanism may not only provide an incomplete assessment of total productivity benefits of exporting but may also jeopardize the measurement of the more traditional direct learning-by-exporting effects because of the endogeneity-inducing omitted variable bias.

In this paper, we extend De Loecker (2013) to develop a unified empirical framework for productivity measurement that explicitly accommodates both the direct LBE channel taking place *within* the firm as well as the indirect LFE channel working *between* firms, which enables us to robustly measure and test these two effects in a consistent structural framework. We do so by formalizing the evolution of firm productivity as an export-controlled process, with the future productivity potentially affected not only by the firm's own export behavior but also by that of its spatially proximate peers in the industry. This allows a simultaneous, "internally consistent" identification of firm productivity and the corresponding effects of exporting. Our identification strategy utilizes the structural link between the parametric production function and the firm's first-order condition for static inputs which helps us circumvent Ackerberg et al.'s (2015) and Gandhi et al.'s (2018) non-identification critiques of conventional proxy-based productivity estimators. In addition, owing to the nonparametric treatment of the firm productivity process, our model enables us to accommodate heterogeneity and nonlinearity in productivity effects of exporting across firms.

We apply our semiparametric methodology to a panel of manufacturing plants in Chile in 1995–2007. We find significant evidence in favor of both the LBE and LFE effects. Overall, the LBE productivity effect is statistically significant for 93% of all firms in our data set, although the results suggest that that the bulk of a productivity boost attributable to (internal) learning from exporting takes place immediately after the domestic firm engages in exports. On average, the size of the LFE effect is comparable to that of LBE. However, at the observation level, the LFE effect is significantly non-zero only for 69% of plants, thus suggesting that the indirect cross-firm productivity-boosting effect of exporting is less prevalent in manufacturing than the direct learning effect taking place within the firm. We also document that less productive plants benefit more from export-driven learning (via both the internal or external channels) thereby suggesting that more productive plants may have less absorptive capacity to learn from their own export experiences as well as to absorb knowledge from their exporting peers. We also find that the more export-oriented firm is, the less (more) it learns by exporting (from exporters). A greater exposure to exporters also helps plants absorb productivity improvements from their own export behavior implying that the two channels are complementary.

Chapter 2

Accounting for Cross-Location Technological Heterogeneity in the Measurement of Operations Efficiency and Productivity[*]

2.1 Introduction

It is well-documented in management, economics as well as operations research that businesses, even in narrowly defined industries, are quite different from one another in terms of productivity. These cross-firm productivity differentials are large, persistent and ubiquitous (see Syverson, 2011). Research on this phenomenon is therefore unsurprisingly vast and includes attempts to explain it from the perspective of firms' heterogeneous behaviors in research and development (e.g., Griffith et al., 2004), corporate operational strategies (Smith and Reece, 1999), ability of the managerial teams (Demerjian et al., 2012), ownership structure (Ehrlich et al., 1994), employee training and education (Moretti, 2004), allocation efficiency (Song et al., 2011), participation in globalization (Grossman and Helpman, 2015) and many others. In most such studies, a common production function/technology is typically assumed for all firms within the industry, and the differences in operations performance of firms are confined to variation in the "total factor productivity," the Solow residual (Solow,

[*] This chapter is based on Malikov et al. (2021).

1957).[1]

In this paper, we approach the heterogeneity in firm performance from a novel perspective in that we explicitly acknowledge the existence of locational effects on the operations technology of firms and their underlying productivity. We allow the firm-level production function to vary across space, thereby accommodating potential neighborhood influences on firm production. In doing so, we are able to examine the role of locational heterogeneity for cross-firm differences in operations performance/efficiency.

A firm's location is important for its operations technology. For example, Ketokivi et al. (2017) show that hospital location is significantly related to its performance and that a hospital's choice of strategy can help moderate the effect of location through the interplay of local environmental factors with organizational strategy. As shown in Figure 2.1, chemical enterprises in China, the focus of empirical analysis in this paper, are widely (and unevenly) distributed across space. Given the sheer size of the country (it is the third largest by area), it is implausible that, even after controlling for firm heterogeneity, all these businesses operate using the same production technology. Organizations in all industries—not only hospitals and chemical manufacturers—develop strategies to respond to local environment and the associated competitive challenges, and those strategies drive operational decisions regarding investments in new or updated technologies.

Theoretically, there are many reasons to believe that the production technology is location-specific. First, exogenous local endowments and institutional environments, such as laws, regulations and local supply chains, play a key role in determining firm performance. The location of firms determines key linkages between the production, market, supply chain and product development (Goldstein et al., 2002). If we look at the global distribution of the supply chains of many products, the product development and design is usually conducted in developed countries such as the U.S. and European countries, while the manufacturing and assembly process is performed in East Asian countries such as China and Vietnam. This spatial distribution largely reflects the endowment differences in factors of production (e.g.,

[1] A few studies alternatively specify an "augmented" production function which, besides the traditional inputs, also admits various firm-specific shifters such as the productivity-modifying factors mentioned above. But such studies continue to assume that the same technology frontier applies to all firms.

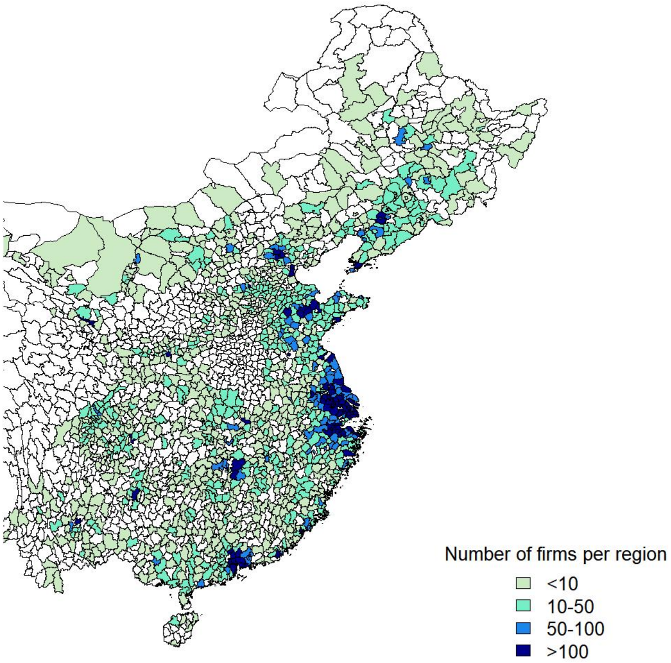

Figure 2.1. Spatial Distribution of Manufacturers of Chemicals in China, 2004–2006

skilled vs. unskilled labor) and the consequent relative input price differentials across countries. Analogously, take the heterogeneity in endowment and institutions across different locations *within* a country. There are many more world's leading universities on the East and West Coasts of the U.S. than in the middle of the country, and they provide thousands of talented graduates each year to the regional development, bolstering growth in flagship industries such as banking and high-tech in those locations. In China, which our empirical application focuses on, networking and political connections are, anecdotally, the key factors for the success of a business in the Northeast regions, whereas the economy on the Southeast Coast is more market-oriented. Furthermore, there are many broadly defined special economics zones (SEZs) in China, which all are characterized by a small designated geographical area, local management, unique benefits and separate customs and administrative procedures (see Crane et al., 2018). According to a report from the China Development Bank, in 2014, there were 6 SEZs, 14 open coastal cities, 4 pilot free-trade areas and 5 financial reform pilot areas. There were also 31 bonded areas, 114 national high-tech development parks, 164 national agricultural technology parks, 85 national eco-industrial parks, 55 national eco-civilization demonstration areas and 283 national modern agriculture demonstration areas. They spread widely in China and support various economic functions, giving rise to locational heterogeneity in the country's production.

Second, most industries are geographically concentrated in general, whereby firms in the same or related industries tend to spatially cluster, benefiting from agglomeration economies reflected, among other things, in their production technologies that bring about localized *aggregate* increasing returns. Ever since Marshall (1920) popularized these ideas, researchers have shown that industry concentration is too great to be explained solely by the differences in exogenous locational factors and that there are at least three behavioral microfoundations for agglomeration: benefits from labor market pooling/sharing, efficiency gains from the collocation of industries with input-output relationships that improves the quality of matches and technology spillovers (see Ellison and Glaeser, 1999; Duranton and Puga, 2004; Ellison et al., 2010; Singh and Marx, 2013). The key idea of agglomeration economies is that geographic proximity reduces the transport costs of goods, people and, perhaps more importantly, ideas. While it is more intuitive that the movement of goods and people is hindered by spatial distance, the empirical evidence from prior studies shows that technology spillovers are also highly localized because knowledge transfers require interaction that proximity facilitates (see Almeida and Kogut, 1999; Alcácer and Chung, 2007; Singh and Marx, 2013). Therefore, owing to the role of local neighborhood influences, firms that produce the same/similar products but are located in regions with different industry concentration levels are expected to enjoy different agglomeration effects on their operations.

Because location is an important factor affecting firm performance, previous empirical studies heavily rely on spatial econometrics to examine the locational/spatial effects on production. Oftentimes, spatially-weighted averages of other firms' outputs and inputs are included as additional regressors in spatial autoregressive (SAR) production-function models (e.g., Glass et al., 2016b, 2020a,b; Vidoli and Canello, 2016; Serpa and Krishnan, 2018; Glass and Kenjegalieva, 2019; Kutlu et al., 2020; Hou et al., 2020). The appropriateness of such a conceptualization of firm-level production functions in the presence of locational influences however remains unclear because these SAR specifications are difficult to reconcile with the theory of firm. For instance, the reduced form of such models effectively implies substitutability of the firm's inputs with those of its peers and does not rule out the possibility of the firm's output increasing when the neighboring firms use more inputs even if the firm itself keeps own inputs fixed and the productivity remains the same. Further, these models continue to implausibly assume that all firms use the same production technology no matter their location. The practical implementation of SAR production-function models is, perhaps, even more problematic: (*i*) they imply additional, highly nonlinear parameter restrictions necessary to ensure that the conventional production axioms are not violated, and (*ii*) they are likely unidentifiable from the data given the inapplicability of available proxy-variable estimators and the pervasive lack of valid external instruments at the firm level. We discuss this in detail in Appendix A.1.[2]

In this paper, we consider a semiparametric production function in which both the input-to-output transformation technology and productivity are location-specific. Concretely, using the location information for firms, we let the input-elasticity and productivity-process parameters be nonparametric functions of the firm's geographic location (latitude and longitude) and estimate these unknown functions via kernel methods. Our methodology captures the cross-firm spatial influences through local smoothing, whereby the production technology for each location is calculated as the geographically weighted average of the input-output *relationships* for firms in the nearby locations with larger weights assigned to the firms that are more spatially proximate. This is fundamentally different from the SAR production-function models that formulate neighborhood influences using spatially-weighed averages of the output/inputs *quantities* while keeping the production technology the same for all firms. Consistent with the agglomeration literature, our approach implies that learning and knowledge spillovers are localized and that their chances/intensity diminish with distance. Importantly, by utilizing the data-driven selection of smoothing parameters that regulate spatial weighting of neighboring firms in kernel smoothing, we avoid the need to rely on *ad hoc* specifications of the weighting schemes and spatial radii of neighborhood influences like the traditional SAR models do. It also allows us to be agnostic about

[2] But we should note that studies of the nexus between location/geography and firm performance in operations research and management are not all confined to the production theory paradigm; e.g., see Bannister and Stolp (1995), Goldstein et al. (2002), Kalnins and Chung (2004, 2006), Dahl and Sorenson (2012) and Kulchina (2016).

the channels through which firm location affects its production, and our methodology inclusively captures all possible mechanisms of agglomeration economies.

Our conceptualization of spatial influences by means of locationally-varying parameters is akin to the idea of “geographically weighted regressions” (GWR) introduced and popularized in the field of geography by Brunsdon et al. (1996); also see Fotheringham et al. (2002) and many references therein. Just like ours, the GWR technique aims to model processes that are not constant over space but exhibit local variations and do so using a varying-coefficient specification estimated via kernel smoothing over locations. However, the principal—and non-trivial—distinction of our methodology from the GWR approach is in its emphasis on *identification* of the spatially varying relationship. Concretely, for consistency and asymptotic unbiasedness the GWR methods rely on the assumption that (non-spatial) regressors in the relationship of interest are mean-orthogonal to the stochastic disturbance which rules out the presence of correlated unobservables as well as the potential simultaneity of regressors and the outcome variable for reasons other than spatial autoregression. The latter two are, however, more the rule rather than the exception for economic relations, which are affected by behavioral choices, including the firm-level production function. Recovering the data generating process underlying the firm’s production operations from observational data (i.e., its identification) requires tackling the correlation between regressors and the error term that the GWR cannot handle, making it unable to consistently estimate the production technology and firm productivity. This is precisely our focus.[3]

The identification of production functions in general, let alone with locational heterogeneity, is not trivial due to the endogeneity issue whereby the firm’s input choices are correlated with its productivity. Complexity stems from the latency of firm productivity. Due to rather unsatisfactory performance of the conventional approaches to identification of production functions, such as fixed effects estimation or instrumentation using prices, there is a growing literature targeted at solving endogeneity using a proxy-variable approach (e.g., see Olley and Pakes, 1996; Levinsohn and Petrin, 2003; Ackerberg et al., 2015; Gandhi et al., 2020) which has gained wide popularity among empiricists.

[3] In effect, our methodology constitutes a generalization of the GWR technique to accommodate endogenous regressors in the context of production-function estimation.

To identify the locationally-varying production functions, we develop a semiparametric proxy-variable estimator that accommodates locational heterogeneity across firms. To this end, we build upon Gandhi et al. (2020) whose framework we extend to incorporate spatial information about the firms in a semiparametric fashion. More specifically, we make use of the structural link between the production function (of the known varying-coefficient functional form) and the optimality condition for a flexible input derived from the firm's static expected profit maximization problem. We propose a two-step estimation procedure and, to approximate the unknown functional coefficients, employ local-constant kernel fitting. Based on the estimated location-specific production functions, we further propose a locational productivity differential decomposition to break down the cross-region production differences that cannot be explained by input usage (i.e., the differential in "total productivity" of firms across locations) into the contributions attributable to differences in available production technologies and to differences in total-factor operations efficiency of firms.

We apply our model to study locationally heterogeneous production technology among Chinese manufacturing firms in the chemical industry in 2002–2004. Based on the results of the data-driven cross-validation as well as formal statistical tests, the empirical evidence provides strong support to the importance and relevance of location for production. Qualitatively, we find that both technology and firm productivity vary significantly across regions. Firms are more likely to exhibit higher (internal) returns to scale in regions of agglomeration. However, the connection between firm productivity and industry concentration across space is unclear. The decomposition analysis reveals that differences in *technology* (as opposed to in idiosyncratic firm heterogeneity) are the main source of cross-location total productivity differentials, on average accounting for 2/3 of the differential.

To summarize, our contribution is as follows. We propose a semiparametric methodology to accommodates locational heterogeneity in the production-function estimation while maintaining the standard structural assumptions about firm production. Unlike the available SAR-type alternatives, our model explicitly estimates the cross-locational variation in production technology. To operationalize our methodology, we extend the widely-used proxy-variable identification methods to incorporate firm location. Our model as well as the proposed decomposition method for disentangling the effects of location on firm productivity from those on technological input-output relationship should provide a valuable addition to the toolkit of empiricists interested in studying agglomeration economies and technology spillovers. In the context of operations management in particular, our methodology will be most useful for empirical studies focused on the analysis of operations efficiency/productivity and its "determinants;" (e.g., Ross and Droge, 2004; Berenguer et al., 2016; Jola-Sanchez et al., 2016; Lam et al., 2016, are just a few recent examples of such analyses). In the case of multi-input production, the "total factor productivity" is among the most popular comprehensive measures of operations efficiency/productivity of the firm, and our paper shows how to measure the latter robustly when production relationships are not constant over space and are subject to neighborhood influences. This is particularly interesting because the effects of location, supply chain integration and agglomeration on firm performance have recently attracted much attention among researchers in operations management (e.g., Goldstein et al., 2002; Ketokivi et al., 2017; Flynn et al., 2010).

The rest of the paper is organized as follows. Section 2.2 describes the model of firm-level production exhibiting locational heterogeneity. We describe our identification and estimation strategy in Section 2.3. We provide the locational productivity differential decomposition in Section 2.4. The empirical application is presented in Section 2.5. Section 3.6 concludes. Supplementary materials are relegated to the Appendix.

2.2 Locational Heterogeneity in Production

Consider the production process of a firm i ($i = 1,\dots,n$) in the time period t ($t = 1,\dots,T$) in which physical capital $K_{it} \in \Re_+$, labor $L_{it} \in \Re_+$ and an intermediate input such as materials $M_{it} \in \Re_+$ are transformed into the output $Y_{it} \in \Re_+$ via a production function given the (unobserved) firm productivity. Also, let S_i be the (fixed) location of firm i, with the obvious choice being $S_i = (\text{lat}_i, \text{long}_i)'$, where lat_i and long_i are the latitude and longitude coordinates of the firm's location. Then, the locationally varying production function is

$$Y_{it} = F_{|S_i}(K_{it}, L_{it}, M_{it}) \exp\{\omega_{it}\} \exp\{\eta_{it}\}, \tag{2.2.1}$$

where $F_{|S_i}(\cdot)$ is the firm's location-specific production function that varies over space (as captured by S_i) to accommodate locational heterogeneity in production technology, ω_{it} is the firm's persistent Hicks-neutral total factor productivity capturing its operations efficiency, and η_{it} is a random transitory shock. Note that, so long as the firm's location is fixed, ω_{it} that persist for the same firm i, by implication, then also has the evolution process that is specific to this firm's location S_i; we expand on this below.

As in Gandhi et al. (2020), Malikov et al. (2020) and Malikov and Zhao (2021), physical capital K_{it} and labor L_{it} are said to be subject to adjustment frictions (e.g., time-to-install, hiring/training costs), and the firm optimizes them dynamically at time $t-1$ rendering these predetermined inputs quasi-fixed at time t. Materials M_{it} is a freely varying (flexible) input and is determined by the firm statically at time t. Thus, both K_{it} and L_{it} are the state variables with dynamic implications and follow their respective deterministic laws of motion:

$$K_{it} = I_{it-1} + (1-\delta)K_{it-1} \quad \text{and} \quad L_{it} = H_{it-1} + L_{it-1}, \tag{2.2.2}$$

where I_{it}, H_{it} and δ are the gross investment, net hiring and the depreciation rate, respectively.

Following the convention, we assume that the risk-neutral firm maximizes a discounted stream of expected life-time profits in perfectly competitive output and factor markets subject to the state variables and expectations about the market structure variables including prices that are common to all firms.[4] Also, for convenience, we denote $\mathscr{I}_{it}$ to be the infor-

[4] We use the perfect competition and homogeneous price assumption mainly for two reasons: (i) it is the most widely used assumption in the literature on structural identification of production function and productivity, and (ii) this assumption has been repeatedly used when studying the same data as ours (e.g., Baltagi et al., 2016; Malikov et al., 2020). Relaxing the perfect competition assumption is possible but non-trivial, and it requires additional assumptions about the output demand (e.g., De Loecker, 2011) and/or extra information on firm-specific output prices that are usually not available for manufacturing data (e.g., De Loecker et al., 2016) or imposing *ex ante* structure on the returns to scale (see Flynn et al., 2019). It is still a subject of ongoing research. Given the emphasis of our contribution on incorporating *technological* heterogeneity (associated with firm location, in our case) in the measurement of firm productivity, we opt to keep all other aspects

mation set available to the firm i for making the period t production decisions.

In line with the proxy variable literature, we model firm productivity ω_{it} as a first-order Markov process which we, however, endogenize à la Doraszelski and Jaumandreu (2013) and De Loecker (2013) by incorporating productivity-enhancing and "learning" activities of the firm. To keep our model as general as possible, we denote all such activities via a generic variable G_{it} which, depending on the empirical application of interest, may measure the firm's R&D expenditures, foreign investments, export status/intensity, etc.[5] Thus, ω_{it} evolves according to a location-inhomogeneous controlled first-order Markov processes with transition probability $\mathscr{P}^{\omega}_{|S_i}(\omega_{it}|\omega_{it-1}, G_{it-1})$. This implies the following location-specific mean regression for firm productivity:

$$\omega_{it} = h_{|S_i}(\omega_{it-1}, G_{it-1}) + \zeta_{it}, \tag{2.2.3}$$

where $h_{|S_i}(\cdot)$ is the location-specific conditional mean function of ω_{it}, and ζ_{it} is a random innovation unanticipated by the firm at period $t-1$ and normalized to zero mean: $\mathbb{E}[\zeta_{it}|\mathscr{I}_{it-1}] = \mathbb{E}[\zeta_{it}] = 0$.

The evolution process in (2.2.3) implicitly assumes that productivity-enhancing activities and learning take place with a delay which is why the dependence of ω_{it} on a control G_{it} is lagged implying that the improvements in firm productivity take a period to materialize. Further, in $\mathbb{E}[\zeta_{it}|\,\mathscr{I}_{it-1}] = 0$ we assume that, due to adjustment costs, firms do not experience changes in their productivity-enhancing investments in light of expected *future* productivity innovations. Since the innovation ζ_{it} represents inherent uncertainty about productivity evolution as well as the uncertainty about the success of productivity-enhancing activities, the firm relies on its knowledge of the *contemporaneous* productivity ω_{it-1} when choosing the level of G_{it-1} in period $t-1$ while being unable to anticipate ζ_{it}. These structural timing assumptions are commonly made in models with controlled productivity processes (e.g., Van Biesebroeck, 2005; Doraszelski and Jaumandreu, 2013, 2018; De Loecker, 2013; Malikov et al., 2020; Malikov and Zhao, 2021) and are needed to identify the within-firm productivity-

of modeling consistent with the convention in the literature to ensure meaningful comparability with most available methodologies.

[5] A scalar variable G_{it} can obviously be replaced with a vector of such variables.

improving learning effects.

We now formalize the firm's optimization problem in line with the above discussion. Under risk neutrality, the firm's optimal choice of freely varying input M_{it} is described by the (static) restricted expected profit-maximization problem subject to the already optimal dynamic choice of quasi-fixed inputs:

$$\max_{M_{it}} P_t^Y F_{|S_i}(K_{it}, L_{it}, M_{it}) \exp\{\omega_{it}\}\theta - P_t^M M_{it}, \tag{2.2.4}$$

where P_t^Y and P_t^M are respectively the output and material prices that, given the perfect competition assumption, need not vary across firms; and $\theta \equiv \mathbb{E}[\exp\{\eta_{it}\}|\ \mathcal{I}_{it}]$. The first-order condition corresponding to this optimization yields the firm's conditional demand for M_{it}.

Building on Doraszelski and Jaumandreu's (2013, 2018) treatment of productivity-enhancing R&D investments (a potential choice of G_{it} in our framework) as a contemporaneous decision, we describe the firm's dynamic optimization problem by the following Bellman equation:

$$\mathbb{V}_t\big(\Xi_{it}\big) = \max_{I_{it}, H_{it}, G_{it}} \Big\{\Pi_{t|S_i}(\Xi_{it}) - \mathrm{C}_t^I(I_{it}) - \mathrm{C}_t^H(H_{it}) - \mathrm{C}_t^G(G_{it}) + \mathbb{E}\Big[\mathbb{V}_{t+1}\big(\Xi_{it+1}\big)\Big|\Xi_{it}, I_{it}, H_{it}, G_{it}\Big]\Big\}, \tag{2.2.5}$$

where $\Xi_{it} = (K_{it}, L_{it}, \omega_{it})' \in \mathcal{I}_{it}$ are the state variables;[6] $\Pi_{t|S_i}(\Xi_{it})$ is the restricted profit function derived as a value function corresponding to the static problem in (2.2.4); and $\mathrm{C}_t^\kappa(\cdot)$ is the cost function for capital ($\kappa = I$), labor ($\kappa = H$) and productivity-enhancing activities ($\kappa = G$).[7] In the above dynamic problem, the level of productivity-enhancing activities G_{it+1} is chosen in time period $t+1$ unlike the amounts of dynamic inputs K_{it+1} and L_{it+1} that are chosen by the firm in time period t (via I_{it} and H_{it}, respectively). Solving (2.2.5) for I_{it}, H_{it} and G_{it} yields their respective optimal policy functions.

An important assumption of our structural model of firm production in the presence of

[6] The firm's location S_i is suppressed in the list of state variables due to its time-invariance.

[7] The assumption of separability of cost functions is unimportant, and one can reformulate (2.2.5) using one $\mathrm{C}_t(I_{it}, H_{it}, G_{it})$ for all dynamic production variables.

locational heterogeneity is that firm location S_i is both fixed and exogenous. However, the identification of locational heterogeneity in production may be complicated by the potentially endogenous spatial sorting problem, whereby more productive firms might *ex ante* sort into the what-then-become high productivity locations. Under this scenario, when we compare firm productivity and technology across locations, we may mistakenly attribute gradients therein to the locational effects such as agglomeration and neighborhood influences, while in actuality it may be merely reflecting the underlying propensity of all firms in a given location to be more productive *a priori*. While there has recently been notable progress in formalizing and understanding these coincident phenomena theoretically (e.g., Behrens et al., 2014; Gaubert, 2018), disentangling firm sorting and spatial agglomeration remains a non-trivial task empirically.[8] However, by including the firm's own lagged productivity in the autoregressive ω_{it} process in (2.2.3), we are able (at least to some extent) to account for this potential self-sorting because sorting into locations is heavily influenced by the firm's own productivity (oftentimes stylized as the "talent" or "efficiency" in theoretical models). That is, the locational heterogeneity in firm productivity and technology in our model is measured after partialling out the contribution of its own past productivity. Incidentally, De Loecker (2013) argues the same in the context of productivity effects of exporting and self-selection of exporters.

2.3 Methodology

This section describes our strategy for (structural) identification and estimation of the firm's location-specific production technology and unobserved productivity.

[8] Urban economics literature also distinguishes the third endogenous process usually referred to as the "selection" which differs from sorting in that it occurs *ex post* after the firms had self-sorted into locations and which determines their continuing survival. We abstract away from this low-productivity-driven attrition issue in the light of the growing empirical evidence suggesting that it explains none of spatial productivity differences which, in contrast, are mainly driven by agglomeration economies (see Combes et al., 2012). Relatedly, the firm attrition out of the sample has also become commonly accepted as a practical non-issue in the productivity literature so long as the data are kept unbalanced. For instance, Levinsohn and Petrin (2003, p.324) write: "The original work by Olley and Pakes devoted significant effort to highlighting the importance of not using an artificially balanced sample (and the selection issues that arise with the balanced sample). They also show once they move to the unbalanced panel, their selection correction does not change their results."

Following the popular practice in the literature (e.g., see Olley and Pakes, 1996; Levinsohn and Petrin, 2003; Doraszelski and Jaumandreu, 2013; Ackerberg et al., 2015; Collard-Wexler and De Loecker, 2015; Konings and Vanormelingen, 2015), we assume the Cobb-Douglas specification for the production function which we adapt to allow for potential locational heterogeneity in production. We do so in a semiparametric fashion as follows:

$$\ln F_{|S_i}(\cdot) = \beta_K(S_i)k_{it} + \beta_L(S_i)l_{it} + \beta_M(S_i)m_{it}, \tag{2.3.1}$$

where the lower-case variables denote the logs of the corresponding upper-case variables, and the input elasticity functions $[\beta_K(\cdot), \beta_L(\cdot), \beta_M(\cdot)]'$ are unspecified smooth functions of the firm's location S_i. The local smoothness feature of the production relationship, including both the input elasticities and persistent productivity process below, captures the effects of technology spillovers and agglomeration economies that give rise to local neighborhood influences. Our methodology can also adopt more flexible specifications such as the log-quadratic translog, which provides a natural extension of the log-linear Cobb-Douglas form. See Appendix A.2 for the details on this extension.

Before proceeding further, we formalize the location-specific autoregressive conditional me-an function of ω_{it} in its evolution process (2.2.3). Following Doraszelski and Jaumandreu (2013, 2019), Ackerberg et al. (2015), Grieco et al. (2016, 2019) and many others, we adopt a parsimonious first-order autoregressive specification of the Markovian evolution for productivity but take a step further by assuming a more flexible semiparametric location-specific formulation akin to that for the production technology in (2.3.1):

$$h_{|S_i}(\cdot) = \rho_0(S_i) + \rho_1(S_i)\omega_{it-1} + \rho_2(S_i)G_{it-1}. \tag{2.3.2}$$

2.3.1 Proxy Variable Identification

Substituting for $F_{|S_i}(\cdot)$ in the locationally varying production function (2.2.1) using (2.3.1), we obtain

$$
\begin{aligned}
y_{it} &= \beta_K(S_i)k_{it} + \beta_L(S_i)l_{it} + \beta_M(S_i)m_{it} + \omega_{it} + \eta_{it} && (2.3.3)\\
&= \beta_K(S_i)k_{it} + \beta_L(S_i)l_{it} + \beta_M(S_i)m_{it} + \rho_0(S_i) + \rho_1(S_i)\omega_{it-1} + \rho_2(S_i)G_{it-1} + \zeta_{it} + \eta_{it}, && (2.3.4)
\end{aligned}
$$

where we have also used the Markov process for ω_{it} from (2.2.3) combined with (2.3.2) in the second line.

Under our structural assumptions, all right-hand-side covariates in (2.3.4) are predetermined and weakly exogenous with respect to $\zeta_{it} + \eta_{it}$, except for the freely varying input m_{it} that the firm chooses in time period t conditional on ω_{it} (among other state variables) thereby making it a function of ζ_{it}. That is, m_{it} is endogenous.

Prior to finding ways to tackle the endogeneity of m_{it}, to consistently estimate (2.3.4), we first need to address the latency of firm productivity ω_{it-1}. A popular solution is a proxy variable approach à la Levinsohn and Petrin (2003) whereby latent productivity is controlled for by inverting the firm's conditional demand for an observable static input such as materials. However, such a standard proxy approach generally fails to identify the firm's production function and productivity due to the lack of a valid instrument (from within the production function) for the endogenous m_{it} despite the abundance of predetermined lags of inputs. As recently shown by Gandhi et al. (2020), identification cannot be achieved using the standard procedure because *no* exogenous higher-order lag provides excluded relevant variation for m_{it} after conditioning the model on the already included self-instrumenting variables. As a result, the production function remains unidentified in flexible inputs. In order to solve this under-identification problem, Gandhi et al. (2020) suggest exploiting a structural link between the production function and the firm's (static) first-order condition for the freely varying input. In what follows, we build on this idea which we modify along the lines of Doraszelski and Jaumandreu (2013) and Malikov and Zhao (2021) in explicitly making use of the assumed functional form of production technology.

First step.—We first focus on the identification of production function in its flexible input m_{it}. Specifically, given the technology specification in (2.3.1), we seek to identify the material elasticity function $\beta_M(S_i)$. To do so, we consider an equation for the firm's first-order condition for the static optimization in (2.2.4). The optimality condition with respect to M_{it} in logs is given by (in logs)

$$\ln P_t^Y + \beta_K(S_i)k_{it} + \beta_L(S_i)l_{it} + \ln\beta_M(S_i) + [\beta_M(S_i)-1]m_{it} + \omega_{it} + \ln\theta = \ln P_t^M, \tag{2.3.5}$$

which can be transformed by subtracting the production function in (2.3.3) from it to obtain the following location-specific material share equation:

$$v_{it} = \ln[\beta_M(S_i)\theta] - \eta_{it}, \tag{2.3.6}$$

where $v_{it} \equiv \ln\left(P_t^M M_{it}\right) - \ln\left(P_t^Y Y_{it}\right)$ is the log nominal share of material costs in total revenue, which is observable in the data.

The material share equation in (2.3.6) is powerful in that it enables us to identify unobservable material elasticity function $\beta_M(S_i)$ using the information about the log material share v_{it}. Specifically, we first identify a "scaled" material elasticity function $\beta_M(S_i)\times\theta$ using the moment condition $\mathbb{E}[\eta_{it}|\mathscr{I}_{it}] = \mathbb{E}\left[\eta_{it}|S_i\right] = 0$, from where we have that

$$\ln[\beta_M(S_i)\theta] = \mathbb{E}[v_{it}|S_i]. \tag{2.3.7}$$

To identify the material elasticity function $\beta_M(S_i)$ net of constant θ, note that θ is

$$\theta \equiv \mathbb{E}\left[\exp\left\{\eta_{it}\right\}\right] = \mathbb{E}\left[\exp\left\{\eta_{it}\right\}\right] = \mathbb{E}\left[\exp\{\mathbb{E}[v_{it}|S_i] - v_{it}\}\right], \tag{2.3.8}$$

which allows us to isolate $\beta_M(S_i)$ via

$$\beta_M(S_i) = \exp\{\mathbb{E}[v_{it}|S_i]\} / \mathbb{E}\left[\exp\{\mathbb{E}[v_{it}|S_i] - v_{it}\}\right]. \tag{2.3.9}$$

By having identified the material elasticity function $\beta_M(S_i)$, we have effectively pinpointed the production technology in the dimension of its endogenous static input thereby effectively circumventing the Gandhi et al. (2020) critique. This is evident when (2.3.4) is rewritten as

$$y^*_{it} = \beta_K(S_i)k_{it} + \beta_L(S_i)l_{it} + \rho_0(S_i) + \rho_1(S_i)\omega_{it-1} + \rho_2(S_i)G_{it-1} + \zeta_{it} + \eta_{it}, \tag{2.3.10}$$

where $y^*_{it} \equiv y_{it} - \beta_M(S_i)m_{it}$ on the left-hand side is already identified/observable and, hence, model in (2.3.10) now contains *no* endogenous regressors that need instrumentation.

Second step.—To identify the rest of the production function, we proxy for latent ω_{it-1} using the known functional form of the conditional material demand function implied by the static first-order condition in (2.3.5) which we analytically invert for productivity. Namely, using the inverted (log) material function $\omega_{it} = \ln[P^M_t/P^Y_t] - \beta_K(S_i)k_{it} - \beta_L(S_i)l_{it} - \ln[\beta_M(S_i)\theta] + [1-\beta_M(S_i)]m_{it}$ to substitute for ω_{it-1} in (2.3.10), we get

$$\begin{aligned} y^*_{it} = {} & \beta_K(S_i)k_{it} + \beta_L(S_i)l_{it} + \rho_0(S_i) + \rho_1(S_i)\Big[v^*_{it-1} - \beta_K(S_i)k_{it-1} - \beta_L(S_i)l_{it-1}\Big] + \\ & \rho_2(S_i)G_{it-1} + \zeta_{it} + \eta_{it}, \end{aligned} \tag{2.3.11}$$

where $v^*_{it-1} = \ln[P^M_{t-1}/P^Y_{t-1}] - \ln[\beta_M(S_i)\theta] + [1-\beta_M(S_i)]m_{it-1}$ is already identified/observable and predetermined with respect to $\zeta_{it} + \eta_{it}$.[9] All regressors in (2.3.11) are weakly exogenous, and this proxied model is identified based on the moment conditions:

$$\mathbb{E}[\zeta_t + \eta_t |\, k_{it}, l_{it}, k_{it-1}, l_{it-1}, G_{it-1}, v^*_{it-1}(m_{it-1}), S_i] = 0. \tag{2.3.12}$$

With the production technology and the transitory shock η_{it} successfully identified in the two previous steps, we can readily recover ω_{it} from (2.3.3) via $\omega_{it} = y_{it} - \beta_K(S_i)k_{it} -$

[9] Following the convention in the literature (e.g., Olley and Pakes, 1996; Levinsohn and Petrin, 2003; Ackerberg et al., 2015; Doraszelski and Jaumandreu, 2013; Gandhi et al., 2020), we assume there is no measurement error in m_{it}. However, if the (log) material input is measured with errors, due to the reasons such as inventories, subcontracting and outsourcing, it will affect both the first- and the second-step estimation. More specifically, adjusting the first step is less problematic if the measurement error is classical since m_{it} is in the dependent variable. However, the second-step equation (2.3.11) will have a new endogeneity issue due to the measurement error. In such a case, additional identifying assumptions are often needed; see Hu et al. (2020) for an example.

$\beta_L(S_i)l_{it} - \beta_M(S_i)m_{it} - \eta_{it}$.

Our identification methodology is also robust to the Ackerberg et al. (2015) critique that focuses on the inability of structural proxy estimators to separably identify the additive production function and productivity proxy. Such an issue normally arises in the wake of perfect functional dependence between freely varying inputs appearing both inside the unknown production function and productivity proxy. Our second-step equation (2.3.11) does not suffer from such a problem because it contains no (endogenous) variable input on the right-hand side, the corresponding elasticity of which has already been identified from the material share equation in the first step.

2.3.2 Semiparametric Estimation

Given the semiparametric varying-coefficient specifications adopted for both the production technology [in (2.3.1)] and productivity evolution [in (2.3.2)], we estimate both the first- and second-step equations (2.3.6) and (2.3.11) via *local* least squares. We employ local-constant kernel fitting.

Denote the unknown $\ln[\beta_M(S_i)\theta]$ as some nonparametric function $b_M(S_i)$. Under the assumption that input elasticity functions are smooth and twice continuously differentiable in the neighborhood of $S_i = s$, unknown $b_M(S_i)$ can be locally approximated around s via $b_M(S_i) \approx b_M(s)$ at points S_i close to s, $|S_i - s| = o(1)$. Therefore, for locations S_i in the neighborhood of s, we can approximate (2.3.6) by

$$v_{it} \approx b_M(s) - \eta_{it}, \tag{2.3.13}$$

with the corresponding local-constant kernel estimator of $\ln[\beta_M(s)\theta]$ given by

$$\widehat{b}_M(s) = \left[\sum_i \sum_t \mathcal{K}_{h_1}(S_i, s)\right]^{-1} \sum_i \sum_t \mathcal{K}_{h_1}(S_i, s)\, v_{it}, \tag{2.3.14}$$

where $\mathcal{K}_{h_1}(S_i, s)$ is a kernel that weights each observation on the basis of proximity of its S_i value to s.

To avoid over-smoothing in dense ranges of the support of the data while under-smoothing in sparse tails, which a “fixed” bandwidth parameter is well-known to produce, we employ an “adaptive” bandwidth capable of adapting to the local distribution of the data. Specifically, to weight observations, we use an h_1-nearest-neighbor bandwidth $R_{h_1}(s)$ defined as the Euclidean distance between the fixed location s and its h_1th nearest location among $\{S_i\}$, i.e.,

$$R_{h_1}(s) = \|S_{(h_1)} - s\|, \tag{2.3.15}$$

where $S_{(h_1)}$ is the h_1th nearest neighbor of s. Evidently, $R_{h_1}(s)$ is just the h_1th order statistic on the distances $\|S_i - s\|$. It is s-specific and, hence, adapts to data distribution. Correspondingly, the kernel weight function is given by

$$\mathscr{K}_{h_1}(S_i, s) = \mathsf{K}\left(\frac{\|S_i - s\|}{R_{h_1}(s)}\right), \tag{2.3.16}$$

where $\mathsf{K}(\cdot)$ is a (non-negative) smooth kernel function such that $\int \mathsf{K}(\|u\|)du = 1$; we use a second-order Gaussian kernel.

The key parameter here that controls the degree of smoothing in the first-step estimator (2.3.14) is the number of nearest neighbors (i.e., locations) h_1, which diverges to ∞ as $n \to \infty$ but slowly: $h_1/n \to 0$. We select the optimal h_1 using the data-driven cross-validation procedure. Also note that, despite the location S_i being multivariate, the parameter h_1 is a scalar because it modulates a univariate quantity, namely the distance. Hence, the bandwidth $R_{h_1}(s)$ is also scalar. That is, unlike in the case of a more standard kernel fitting based on fixed bandwidths when the data are weighted using the product of univariate kernels corresponding to each element in $S_i - s$, the adaptive kernel fitting weights data using a *norm* of the vector $S_i - s$. For this reason, when employing nearest neighbor methods, the elements of smoothing variables are typically rescaled so that they are all comparable because, when S_i is multivariate, the nearest neighbor ordering is not scale-invariant. In our case however, we do *not* rescale the elements of S_i (i.e., latitude and longitude) because they are already measured on the same scale and the (partial) distances therein have a concrete physical interpretation.

From (2.3.9), the first-step estimator of $\beta_M(s)$ is

$$\widehat{\beta}_M(s) = nT\exp\left\{\widehat{b}_M(s)\right\}\Big/\sum_i\sum_t \exp\left\{\widehat{b}_M(s) - v_{it}\right\}. \tag{2.3.17}$$

We construct $\widehat{y}^*_{it} \equiv y_{it} - \widehat{\beta}_M(S_i)m_{it}$ and $\widehat{v}^*_{it-1} = \ln[P^M_{t-1}/P^Y_{t-1}] - \ln[\widehat{\beta}_M(S_i)\theta] + [1-\widehat{\beta}_M(S_i)]m_{it-1}$ using the first-step local estimates of $\beta_M(S_i)$. Analogous to the first-step estimation, we then locally approximate each unknown parameter function in (2.3.11) around $S_i = s$ via local-constant approach. Therefore, for locations S_i near s, we have

$$\widehat{y}^*_{it} \approx \beta_K(s)k_{it} + \beta_L(s)l_{it} + \rho_0(s) + \rho_1(s)\left[\widehat{v}^*_{it-1} - \beta_K(s)k_{it-1} - \beta_L(s)l_{it-1}\right] + \rho_2(s)G_{it-1} + \zeta_{it} + \eta_{it}. \tag{2.3.18}$$

Denoting all unknown parameters in (2.3.18) collectively as $\Theta(s) = [\beta_K(s), \beta_L(s), \rho_0(s), \rho_1(s), \rho_2(s)]'$, we estimate the second-step equation via locally weighted nonlinear least squares. The corresponding kernel estimator is

$$\widehat{\Theta}(s) = \arg\min_{\Theta(s)} \sum_i\sum_t \mathscr{K}_{h_2}(S_i, s)\Big(\widehat{y}^*_{it} - \beta_K(s)k_{it} - \beta_L(s)l_{it} - \rho_0(s) - \rho_1(s)\left[\widehat{v}^*_{it-1} - \beta_K(s)k_{it-1} - \beta_L(s)l_{it-1}\right] + \rho_2(s)G_{it-1}\Big)^2, \tag{2.3.19}$$

where h_2 is the number of nearest neighbors of a fixed location s in the second-step estimation. It diverges faster than does the first-step smoothing parameter h_1 so that the first-step estimation has an asymptotically ignorable impact on the second step.

Lastly, the firm productivity is estimated as $\widehat{\omega}_{it} = y_{it} - \widehat{\beta}_K(S_i)k_{it} - \widehat{\beta}_L(S_i)l_{it} - \widehat{\beta}_M(S_i)m_{it} - \widehat{\eta}_{it}$ using the results from both steps.

Finite-Sample Performance. Before applying our proposed methodology to the data, we first study its performance in a small set of Monte Carlo simulations. The results are encouraging, and simulation experiments show that our estimator recovers the true parameters well. As expected of a consistent estimator, the estimation becomes more stable as the sample size grows. For details, see Appendix A.3.

Inference. Due to a multi-step nature of our estimator as well as the presence of nonparametric components, computation of the asymptotic variance of the estimators is not simple. For statistical inference, we therefore use bootstrap. We approximate sampling distributions of the estimators via wild residual block bootstrap that takes into account a panel structure of the data, with all the steps bootstrapped jointly owing to a sequential nature of our estimation procedure. The bootstrap algorithm is described in Appendix A.4.

Testing of Location Invariance. Given that our semiparametric locationally varying production model nests a more traditional fixed-parameter specification that implies locational invariance of the production function and the productivity evolution as a special case, we can formally discriminate between the two models to see if the data support our more flexible modeling approach. We discuss this specification test in detail in Appendix A.5.

2.4 Locational Productivity Differential Decomposition

Since the production function can vary across space, a meaningful comparison of productivity for firms dispersed across space now requires that locational differences in technology be explicitly controlled. That is, the productivity differential between two firms is no longer limited to the difference in their firm-specific total factor productivities ω_{it} (unless they both belong to the same location) because either one of the firms may have access to a more productive technology $F_S(\cdot)$. Given that locational heterogeneity in production is the principal focus of our paper, in what follows, we provide a procedure for measuring and decomposing firm productivity differentials across any two locations of choice.

Let $\mathscr{L}(s,t)$ represent a set of n_t^s firms operating in location s in the year t. For each of these firms, the estimated Cobb-Douglas production function (net of random shocks) in logs is

$$\widehat{y}_{it}^{s} = \widehat{\beta}_K(s)k_{it}^{s} + \widehat{\beta}_L(s)l_{it}^{s} + \widehat{\beta}_M(s)m_{it}^{s} + \widehat{\omega}_{it}^{s}, \tag{2.4.1}$$

where we have also explicitly indexed these firms' observable output/inputs as well as the estimated productivities using the location. Averaging over these firms, we arrive at the "mean" production function for location s in time t:

$$\overline{y}_t^s = \widehat{\beta}_K(s)\overline{k}_t^s + \widehat{\beta}_L(s)\overline{l}_t^s + \widehat{\beta}_M(s)\overline{m}_t^s + \overline{\omega}_t^s, \tag{2.4.2}$$

where $\overline{y}_t^s = \frac{1}{n_t^s}\sum_i \widehat{y}_{it}^s \mathbb{1}\{i \in \mathscr{L}(s,t)\}$, $\overline{x}_t^s = \frac{1}{n_t^s}\sum_i x_{it}^s \mathbb{1}\{i \in \mathscr{L}(s,t)\}$ for $x \in \{k,l,m\}$, and $\overline{\omega}_t^s = \frac{1}{n_t^s}\sum_i \widehat{\omega}_{it}^s$ $\mathbb{1}\{i \in \mathscr{L}(s,t)\}$.

Taking the difference between (2.4.2) and the analogous mean production function for the benchmark location of interest κ in the same year, we obtain the mean *output* differential between these two locations (in logs):

$$\underbrace{\overline{y}_t^s - \overline{y}_t^\kappa}_{\Delta\overline{y}_t^{s,\kappa}} = \left[\widehat{\beta}_K(s)\overline{k}_t^s + \widehat{\beta}_L(s)\overline{l}_t^s + \widehat{\beta}_M(s)\overline{m}_t^s\right] - \left[\widehat{\beta}_K(\kappa)\overline{k}_t^\kappa + \widehat{\beta}_L(\kappa)\overline{l}_t^\kappa + \widehat{\beta}_M(\kappa)\overline{m}_t^\kappa\right] + \left[\overline{\omega}_t^s - \overline{\omega}_t^\kappa\right]. \tag{2.4.3}$$

To derive the mean *productivity* differential (net of input differences) between these two locations, we add and subtract the s location's production technology evaluated at the κ location's inputs, i.e., $\left[\widehat{\beta}_K(s)\overline{k}_t^\kappa + \widehat{\beta}_L(s)\overline{l}_t^\kappa + \widehat{\beta}_M(s)\overline{m}_t^\kappa\right]$, in (2.4.3):

$$\begin{aligned}\Delta\overline{\mathrm{PROD}}_t^{s,\kappa} &\equiv \Delta\overline{y}_t^{s,\kappa} - \widehat{\beta}_K(s)\Delta\overline{k}_t^{s,\kappa} - \widehat{\beta}_L(s)\Delta\overline{l}_t^{s,\kappa} - \widehat{\beta}_M(s)\Delta\overline{m}_t^{s,\kappa} \\ &= \underbrace{\left[\widehat{\beta}_K(s) - \widehat{\beta}_K(\kappa)\right]\overline{k}_t^\kappa + \left[\widehat{\beta}_L(s) - \widehat{\beta}_L(\kappa)\right]\overline{l}_t^\kappa + \left[\widehat{\beta}_M(s) - \widehat{\beta}_M(\kappa)\right]\overline{m}_t^\kappa}_{\Delta\overline{\mathrm{TECH}}_t^{s,\kappa}} + \underbrace{\left[\overline{\omega}_t^s - \overline{\omega}_t^\kappa\right]}_{\Delta\overline{\mathrm{TFP}}_t^{s,\kappa}},\end{aligned} \tag{2.4.4}$$

where $\Delta\overline{x}_t^{s,\kappa} = \overline{x}_t^s - \overline{x}_t^\kappa$ for $x \in \{k,l,m\}$.

Equation (2.4.4) measures mean productivity differential across space and provides a *counterfactual* decomposition thereof. By utilizing the counterfactual output that, given its location-specific technology, the average firm in location s would have produced using the mean inputs employed by the firms in location κ in year t $[\widehat{\beta}_K(s)\overline{k}_t^\kappa + \widehat{\beta}_L(s)\overline{l}_t^\kappa + \widehat{\beta}_M(s)\overline{m}_t^\kappa]$, we are able to measure the locational differential in the mean productivity of firms in the locations s and κ that is *un*explained by their different input usage: $\Delta\overline{\mathrm{PROD}}_t^{s,\kappa}$. More importantly, we can then decompose this locational differential in the total productivity into the contribution attributable to the difference in production technologies $\Delta\overline{\mathrm{TECH}}_t^{s,\kappa}$ and to the difference in the average total-factor operations efficiencies $\Delta\overline{\text{TFP}}_t^{s,\kappa}$.

The locational productivity differential decomposition in (2.4.4) is time-varying, but should one be interested in a scalar measure of locational heterogeneity for the entire sample period, time-specific averages can be replaced with the “grand” averages computed by pooling over all time periods.

2.5 Empirical Application

Using our proposed model and estimation methodology, we explore the locationally heterogeneous production technology among manufacturers in the Chinese chemical industry. We report location-specific elasticity and productivity estimates for these firms and then decompose differences in their productivity across space to study if the latter are mainly driven by the use of different production technologies or the underlying total factor productivity differentials.

2.5.1 Data

We use the data from Baltagi et al. (2016). The dataset is a panel of $n = 12{,}490$ manufacturers of chemicals continuously observed over the 2004–2006 period ($T = 3$). The industry includes manufacturing of basic chemical materials (inorganic acids and bases, inorganic salts and organic raw chemical materials), fertilizers, pesticides, paints, coatings and adhesives, synthetic materials (plastic, synthetic resin and fiber) as well as daily chemical products (soap and cleaning compounds). The original source of these firm-level data is the Chinese Industrial Enterprises Database survey conducted by China’s National Bureau of Statistics (NBS) which covers all state-owned firms and all non-state-owned firms with sales above 5 million Yuan (about \$0.6 million). Baltagi et al. (2016) have geocoded the location of each firm at the zipcode level in terms of the longitude and latitude (the “S” variables) using their postcode information in the dataset. The coordinates are constructed for the location of each firm’s headquarters and are time-invariant. By focusing on the continually operating firms, we mitigate a potential impact of spatial sorting (as well as the attrition due

Table 2.1. Data Summary Statistics

Variables	Mean	1st Qu.	Median	3rd Qu.
—Production Function Variables—				
Output	86,381.98	11,021.09	23,489.53	59,483.71
Capital	35,882.40	1,951.47	5,319.28	17,431.35
Labor	199.07	43.00	80.00	178.00
Materials	48,487.82	5,896.49	12,798.35	33,063.81
—Productivity Controls—				
Skilled Labor Share	0.174	0.042	0.111	0.242
Foreign Equity Share	0.140	0.000	0.000	0.000
Exporter	0.237			
State-Owned	0.051			
—Location Variables—				
Longitude	2.041	1.984	2.068	2.102
Latitude	0.557	0.504	0.547	0.632

Output, capital and materials are in 1,000s of 1998 RMB. Labor is measured in the number of employees. The skilled labor share and foreign equity share are unit-free proportions. The exporter and state-owned variables are binary indicators. The location coordinates are in radians.

to non-survival) on the estimation results and treat the firm location as fixed (exogenous). The total number of observations is 37,470.

Figure 2.1 shows the spatial distribution of firms in our dataset on a map of mainland China (we omit area in the West with no data in our sample). The majority are located on the East Coast and in the Southeast of China, especially around the Yangtze River Delta that is generally comprised of Shanghai and the surrounding areas, the southern Jiangsu province and the northern Zhejiang province.

The key production variables are defined as follows. Output (Y) is measured using sales. The labor input (L) is measured by the number of employees. Capital stock (K) is the net fixed assets for production and operation, and the materials (M) are defined as the expenditure on direct materials. Output, capital and materials are deflated to the 1998 values using the producer price index, the price index for investment in fixed assets and the purchasing price index for industrial inputs, respectively, where the price indices are obtained from the NBS. The unit of monetary values is thousands RMB (Chinese Yuan).

We include four productivity-modifying variables in the evolution process of firm productivity ω_{it}: the share of high-skilled workers (G_1), which is defined as the fraction of workers with a university or comparable education and is time-invariant because the data on workers' education level are only available for 2004; the foreign equity share (G_2), which is measured by the proportion of equity provided by foreign investors; a binary export status indicator (G_3), which takes value one if the firm is an exporter and zero otherwise; and a binary state/public ownership status indicator (G_4), which takes value one if the firm is state-owned and zero otherwise.

Table 3.1 shows the summary statistics, including the mean, 1st quartile, median and 3rd quartile for the variables. For the production-function variables, the mean values are significantly larger than their medians, which suggests that their distributions are skewed to the right. Among firms in the chemical industry, 23.7% are exporters and 5.1% are state-owned. Most firms do not have foreign investors, and the average ratio of foreign to total equity is 0.14. On average, 17.4% of employees in the industry have a college degree or equivalent.

2.5.2 Estimation Results

In order to estimate the locationally varying (semiparametric) production function and firm productivity process in (2.3.1)–(2.3.2), we use the data-driven leave-one-location-out cross-validation method to choose the optimal number of nearest neighboring locations in each step of the estimation (h_1 and h_2) to smooth over "contextual" location variables S_i inside the unknown functional coefficients. This smoothing parameters regulate spatial weighting of neighboring firms in kernel fitting and, as noted earlier, by selecting it via a data-driven procedure, we avoid the need to rely on *ad hoc* specifications of both the spatial weights and radii defining the extent of neighborhood influences. The optimal h_1 and h_2 values are 520 and 340 firm-years in the first- and second-step estimation, respectively. On average across all s, the corresponding adaptive bandwidths are 0.0171 and 0.0169 radians. These bandwidth values are reasonable, given our sample size and the standard deviations of the longitude and latitude,[10] and, evidently, are *not* too large to "smooth out" the firm location to imply location invariance/homogeneity. In fact, we can argue this more formally if kernel-smoothing is done using *fixed* bandwidths so that we can rely on the theoretical results by

[10] For the reference, the sample standard deviations for the longitude and latitude are respectively 0.0889 and 0.0941 radians.

Hall et al. (2007), whereby local-constant kernel methods can remove irrelevant regressors via data-driven over-smoothing (i.e., by selecting large bandwidths). When we re-estimate our locationally-varying model in this manner, the optimal fixed bandwidths for the longitude and latitude in the first-step estimation are 0.009 and 0.010 radians, respectively; the corresponding second-step bandwidths are 0.024 and 0.023 radians. Just like in the case of adaptive bandwidths, these bandwidth values are fairly small relative to variation in the data, providing strong evidence in support of the overall relevancy of geographic location for firm production (i.e., against location invariance). Our location-varying formulation of the production technology and productivity is also formally supported by the Ullah (1985) specification test described in Appendix A.5. Using cross-validated fixed bandwidths, the bootstrap p-value is 0.001. At the conventional significance level, our locationally heterogeneous production model is confidently preferred to a location-invariant formulation.

In what follows, we discuss our semiparametric results obtained using adaptive bandwidths. For inference, we use the bias-corrected bootstrap percentile intervals as described in Appendix A.4. The number of bootstrap replications is set to $B = 1,000$.

Production Function. We first report production-function estimates from our main model in which the production technology is locationally heterogeneous. We then compare these estimates with those obtained from the more conventional, location-invariant model that *a priori* assumes common production technology for all firms. The latter “global” formulation of the production function postulates constancy of the production relationship over space. This model is therefore fully parametric (with constant coefficients) and a special case of our locationally-varying model when S_i is fixed across all i. Its estimation is straightforward and follows directly from (2.3.14)–(2.3.18) by letting the adaptive bandwidths in both steps diverge to ∞ which, in effect, obviates the need to locally weight the data because all kernels will be the same (for details, see Appendix A.5).[11]

Since our model has location-specific input elasticities, there is a distribution of them

[11] Following a suggestion provided by a referee, we also estimate the location-invariant model with location fixed effects added to the production function during the estimation. We find the results do not change much from including these location effects, and therefore these results are not reported.

Table 2.2. Input Elasticity Estimates

	Locationally Varying				*Location-Invariant*
	Mean	1st Qu.	Median	3rd Qu.	Point Estimate
Capital	0.112	0.095	0.115	0.128	0.130
	(0.104, 0.130)	(0.083, 0.116)	(0.110, 0.130)	(0.119, 0.147)	(0.118, 0.141)
Labor	0.303	0.272	0.293	0.342	0.299
	(0.285, 0.308)	(0.248, 0.284)	(0.278, 0.293)	(0.313, 0.356)	(0.280, 0.318)
Materials	0.480	0.452	0.481	0.503	0.495
	(0.466, 0.501)	(0.414, 0.467)	(0.437, 0.502)	(0.456, 0.524)	(0.460, 0.519)

The left panel summarizes point estimates of $\beta_\kappa(S_i)\ \forall\ \kappa \in \{K, L, M\}$ with the corresponding two-sided 95% bias-corrected confidence intervals in parentheses. The right panel reports their counterparts from a fixed-coefficient location-invariant model.

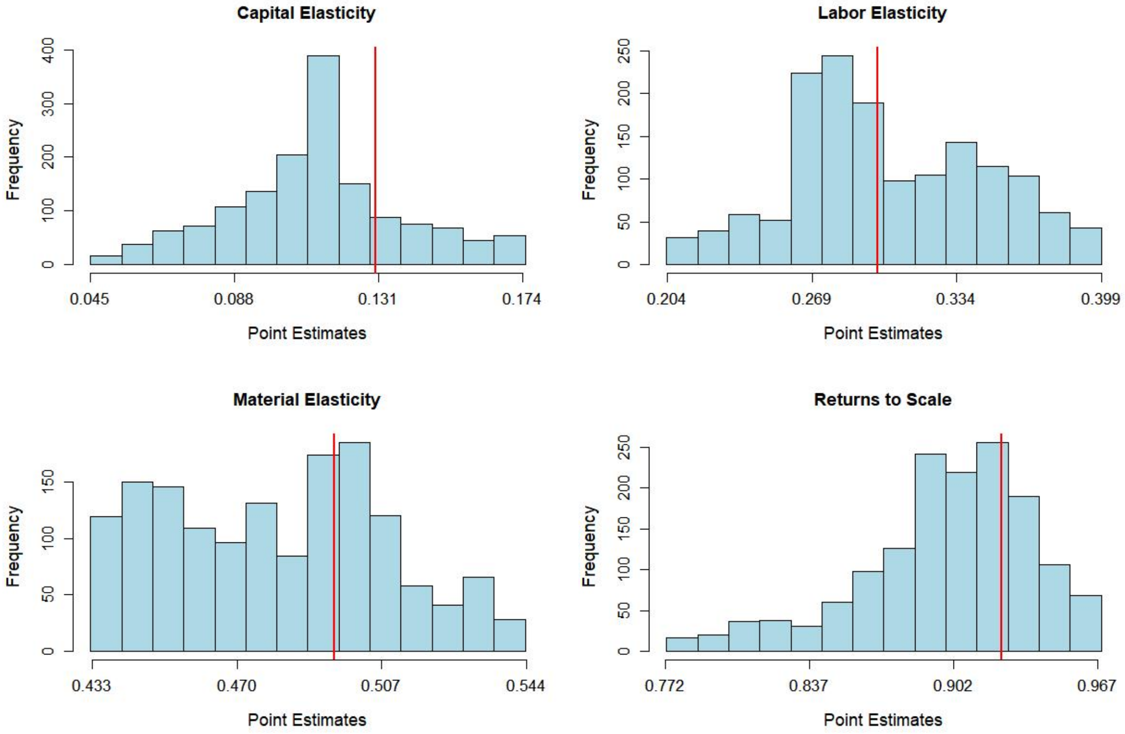

Figure 2.2. Input Elasticity Estimates
(Notes: Vertical lines correspond to location-invariant estimates)

Table 2.3. Locationally Varying Returns to Scale Estimates

	Mean	1st Qu.	Median	3rd Qu.	= 1	< 1
RTS	0.895 (0.820, 0.931)	0.875 (0.801, 0.912)	0.903 (0.827, 0.942)	0.929 (0.855, 0.968)	21.6%	82.3%

The left panel summarizes point estimates of $\sum_{\kappa} \beta_{\kappa}(S_i)$ with $\kappa \in \{K, L, M\}$ with the corresponding two-sided 95% bias-corrected confidence intervals in parentheses. The counterpart estimate of the returns to scale from a fixed-coefficient location-invariant model is 0.924 (0.865, 0.969). The right panel reports the shares of locations in which location-specific point estimates are (*i*) not significantly different from 1 (constant returns to scale) and (*ii*) statistically less than 1 (decreasing returns to scale). The former classification is based on a two-sided test, the latter is on a one-sided test.

(over space) and Table 2.2 summarizes their point estimates. The table also reports the elasticity estimates from the alternative, location-invariant model. The corresponding two-sided 95% bias-corrected confidence intervals for these statistics are reported in parentheses. Based on our model, the mean (median) capital, labor and material elasticity estimates are 0.112, 0.303 and 0.480 (0.115, 0.293 and 0.481), respectively. Importantly, these location-specific elasticities show significant variation. For the capital and labor inputs, the first quartiles are significantly different from the third quartiles. Within the inter-quartile interval of their point estimates, elasticities of capital, labor and materials respectively increase by 0.033, 0.070 and 0.051, which in turn correspond to the 35%, 26% and 11% changes.

In comparison, the elasticity estimates from the location-invariant production function with fixed coefficients are all larger than the corresponding median estimates from our model and fall in between the second and third quartiles of our locationally-varying point estimates. Figure 2.2 provides visualization of the non-negligible technological heterogeneity in the chemicals production technology across different locations in China, which the traditional location-invariant model assumes away. The figure plots histograms of the estimated location-specific input elasticities (and the returns to scale) with the location-invariant counterpart estimates depicted by vertical lines. Consistent with the results in Table 2.2, all distributions show relatively wide dispersion, and the locationally homogeneous model is apparently unable to provide a reasonable representation of production technology across different regions.

Table 2.3 provides summary statistics of the estimated returns to scale (RTS) from our locationally varying production function (also see the bottom-right plot in Figure 2.2). The

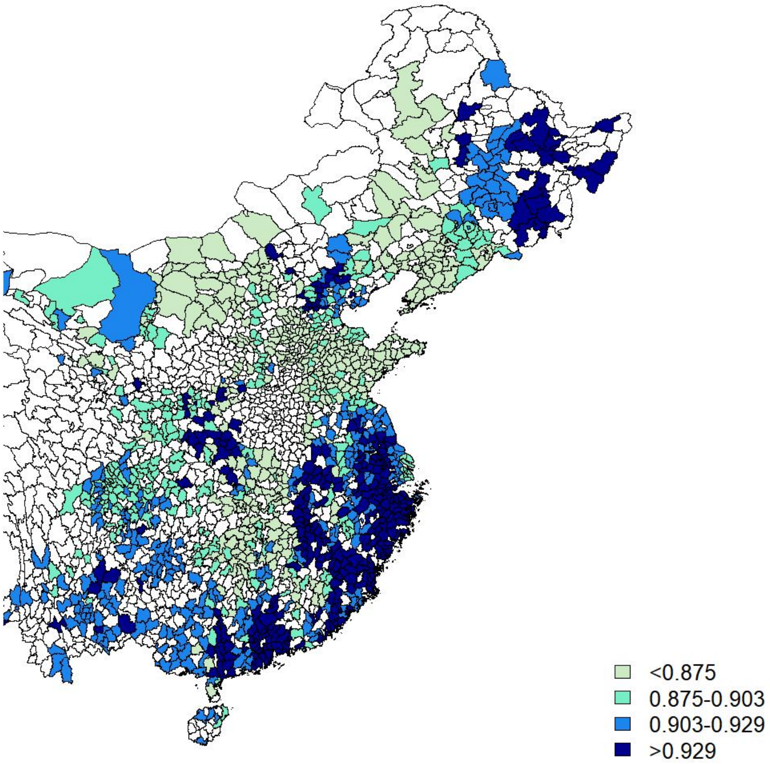

Figure 2.3. Spatial Distribution of Returns to Scale Estimates
(Notes: The color shade cutoffs correspond to the first, second (median) and third quartiles)

mean RTS is 0.895, and the median is 0.903, with the inter-quartile range being 0.054. The right panel of Table 2.3 reports the fraction of locations in which the Chinese manufacturers of chemicals exhibit constant or decreasing returns to scale. This classification is based on the RTS point estimate being statistically equal to or less than one, respectively, at the 5% significance level. The "= 1" classification is based on a two-sided test, whereas the "< 1" test is one-sided. In most locations in China (82.3%), the production technologies of the chemicals firms exhibit *dis*economies of scale, but 21.6% regions show evidence of the constant returns to scale (i.e., scale efficiency).

To further explore the locational heterogeneity in the production technology for chemicals in China, we plot the spatial distribution of the RTS estimates in the country in Figure 2.3. We find that the firms with the largest RTS are mainly located in the Southeast Coast provinces and some parts of the West and Northeast China. The area nearby Beijing also exhibits larger RTS. There are a few possible explanations of such a geographic distribution of the returns to scale. As noted earlier, spillovers and agglomeration have positive effects on the marginal productivity of inputs which typically take form of the scale effects, and they may explain the high RTS on the Southeast Coast and in the Beijing area. Locality-specific resources, culture and polices can also facilitate firms' production process. For example, the rich endowment of the raw materials like coal, phosphate rock and sulfur make the provinces such as Guizhou, Yunnan and Qinghai among the largest fertilizer production zones in China. Furthermore, RTS is also related to the life cycle of a firm. Usually, it is the small, young and fast-growth firms that enjoy higher RTS, whereas the more mature firms that have grown bigger will have transitioned to the low-RTS scale. This may explain the prevalence of the higher-RTS firms in the West and Northeast China.

Productivity Process. We now analyze our semiparametric estimates of the firm productivity process in (2.3.2). Table 2.4 summarizes point estimates of the location-specific marginal effects of productivity determinants in the evolution process of ω_{it}, with the corresponding two-sided 95% bias-corrected confidence intervals in parentheses. In the last column of the left panel, for each productivity-enhancing control G_{it}, we also report the share of locations in which location-specific point estimates are statistically positive (at a 5% significance level) as inferred via a one-sided test.

The autoregressive coefficient on the lagged productivity, which measures the persistence of ω_{it}, is 0.576 at the mean and 0.597 at the median, with the quartile statistics varying from 0.518 to 0.641. It is significantly positive for firms in virtually all locations. For firms in most locations (85.7%), skilled labor has a large and significantly positive effect on productivity: a percentage point increase in the skilled labor share is associated with an improvement in the next period's firm productivity by about 0.4%, on average. Point estimates of the foreign ownership effect are positive in the majority of locations, but firms in only about half the locations benefit from a statistically positive productivity-boosting effect of the inbound foreign direct investment, with the average magnitude of only 7/50 of that attributable to hiring more skilled labor. In line with the empirical evidence reported for China's manufacturing in the literature (see Malikov et al., 2020, and references therein), firms in most regions show insignificant (and negative) effects of the export status on productivity. The

Table 2.4. Productivity Process Coefficient Estimates

	Locationally Varying					*Location-Invariant*
Variables	Mean	1st Qu.	Median	3rd Qu.	> 0	Point Estimate
Lagged Productivity	0.576	0.518	0.597	0.641	99.9%	0.497
	(0.540, 0.591)	(0.469, 0.541)	(0.553, 0.614)	(0.580, 0.665)		(0.455, 0.530)
Skilled Labor Share	0.387	0.287	0.419	0.500	85.7%	0.387
	(0.346, 0.395)	(0.241, 0.309)	(0.345, 0.459)	(0.471, 0.493)		(0.345, 0.425)
Foreign Equity Share	0.054	–0.001	0.062	0.103	47.7%	0.056
	(0.006, 0.074)	(–0.034, 0.066)	(0.033, 0.069)	(0.099, 0.099)		(0.036, 0.075)
Exporter	–0.001	–0.032	–0.005	0.038	24.0%	0.006
	(–0.011, 0.018)	(–0.041, –0.016)	(–0.012, 0.013)	(0.025, 0.067)		(–0.008, 0.018)
State-Owned	0.005	–0.052	0.007	0.073	29.6%	–0.043
	(–0.021, 0.010)	(–0.101, –0.014)	(–0.028, 0.025)	(0.062, 0.076)		(–0.072, –0.009)

The left panel summarizes point estimates of $\rho_j(S_i)\ \forall\ j = 1, \ldots, \dim(G)$ with the corresponding two-sided 95% bias-corrected confidence intervals in parentheses. Reported is also a share of locations in which location-specific point estimates are statistically positive as inferred via a one-sided test. The right panel reports the counterparts from a fixed-coefficient location-invariant model.

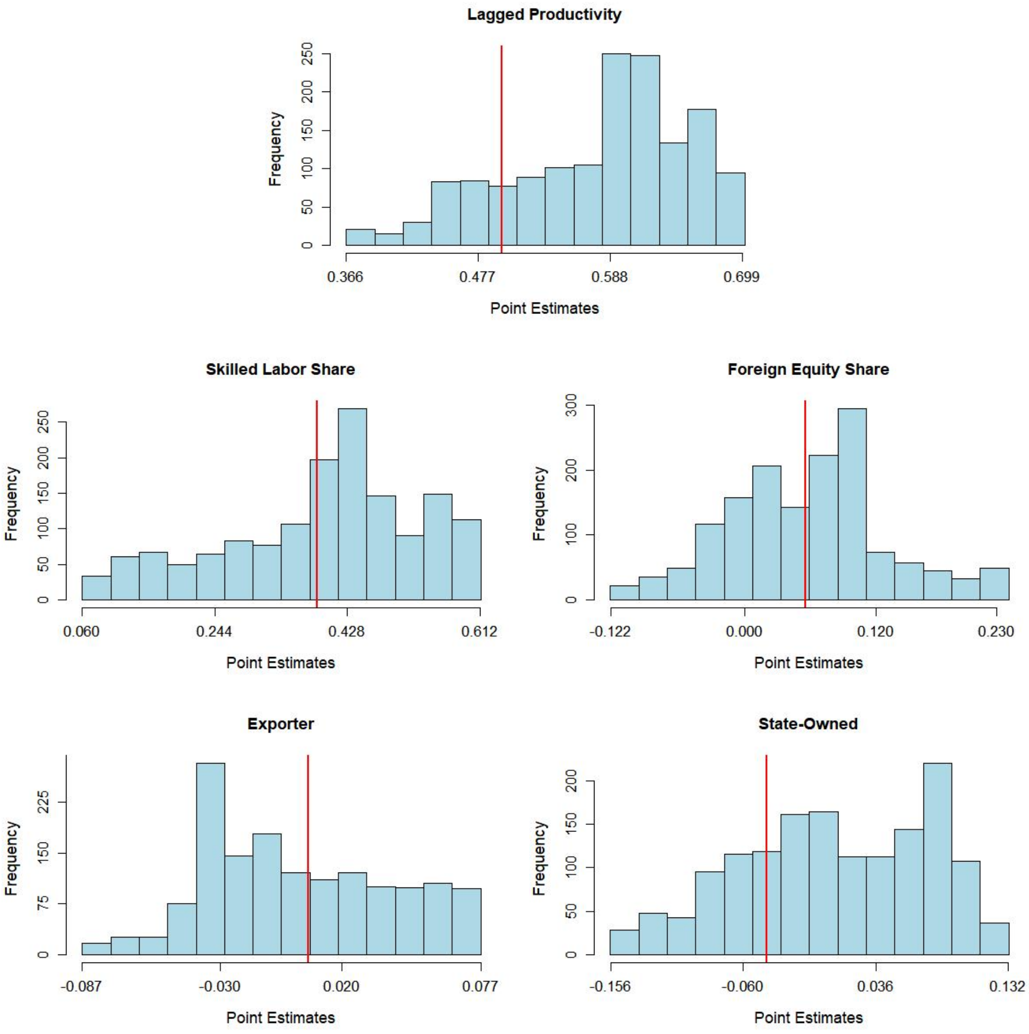

Figure 2.4. Productivity Process Coefficient Estimates
(Notes: Vertical lines correspond to location-invariant estimates)

"learning by exporting" effects are very limited and statistically positive in a quarter of locations only. Interestingly, we find that state/public ownership is a significantly positive contributor to the improvements in firm productivity in about a third of the locations in which the Chinese chemicals manufacturing firms operate. This may be because the less productive state firms exited the market during the market-oriented transition in the late 1990s and early 2000s (prior to our sample period), and the remaining state-owned firms are larger and more productive (also see Hsieh and Song, 2015; Zhao et al., 2020). Another potential reason may be that state ownership could have brought non-trivial financing benefits to these firms which, otherwise, were usually financially constrained due to the under-developed financial market in China during that period.

The far right panel of Table 2.4 reports productivity effects of the G_{it} controls estimated using the location-invariant model. Note that, under the assumption of a location-invariant production, the evolution process of ω_{it} becomes a parametric linear model, and there is only one point estimate of each fixed marginal effect for all firms. Comparing these estimates with the median estimates from our model, the location-invariant marginal effects tend to be smaller. While the persistence coefficient as well as fixed coefficients on the skilled labor and foreign equity shares are positive and statistically significant, the location-invariant estimate of the state ownership effect on productivity is however significantly negative (for all firms, by design). Together with the tendency of a location-invariant model to underestimate, this underscores the importance of allowing sufficient flexibility in modeling heterogeneity across firms (across different locations, in our case) besides the usual Hicks-neutral TFP.

The contrast between the two models is even more apparent in Figure 2.4, which plots the distributions of estimated marginal effects of the productivity-enhancing controls. Like before, the location-invariant counterparts are depicted by vertical lines. The distribution of each productivity modifier spans a relatively wide range, and the corresponding location-invariant estimates are evidently not good representatives for the centrality of these distributions. For example, the productivity-boosting effect of the firm's skilled labor roughly varies between 0.06 and 0.61% per unit percentage point increase in the skilled labor share,

Table 2.5. Locational Productivity Differential Decomposition

Components	Mean	1st Qu.	Median	3rd Qu.
	Locationally Varying Model			
$\Delta\overline{\text{TECH}}^{s,\kappa}$	1.292	1.135	1.331	1.521
$\Delta\overline{\text{TFP}}^{s,\kappa}$	0.574	0.203	0.571	0.893
$\Delta\overline{\text{PROD}}^{s,\kappa}$	1.866	1.652	1.869	2.086
	Location-Invariant Model			
$\Delta\overline{\text{PROD}}^{s,\kappa}$	1.797	1.589	1.816	2.040

The top panel summarizes point estimates of the locational mean productivity differential $\Delta\overline{\text{PROD}}^{s,\kappa} = \Delta\overline{\text{TECH}}^{s,\kappa} + \Delta\overline{\text{TFP}}^{s,\kappa}$ with the corresponding two-sided 95% bias-corrected confidence intervals in parentheses. The bottom panel reports the counterparts from a fixed-coefficient location-invariant model for which, by construction, $\Delta\overline{\text{PROD}}^{s,\kappa} = \Delta\overline{\text{TFP}}^{s,\kappa}$ with $\Delta\overline{\text{TECH}}^{s,\kappa} = 0$. In both cases, the decomposition is pooled for the entire sample period and the benchmark/reference location κ is the one with the smallest mean production: $\kappa = \arg\min_s \overline{y}^s$.

depending on the location. The distribution of this marginal effect across locations is somewhat left-skewed, and the corresponding location-invariant effect estimate evidently does not measure central tendency of these locationally-varying effects well. Similar observations can be made about other varying coefficients in the productivity process.

Productivity Decomposition. We now examine the average productivity differentials for firms in different regions. To this end, we perform the locational decomposition proposed in Section 2.4 to identify the sources of production differences that cannot be explained by input usage. Recall that, by our decomposition, the locational differential in the mean total productivity ($\Delta\overline{\text{PROD}}_t^{s,\kappa}$) accounts for the cross-regional variation in both the input elasticities ($\Delta\overline{\text{TECH}}_t^{s,\kappa}$) and the total factor productivity ($\Delta\overline{\text{TFP}}^{s,\kappa}$). It is therefore more inclusive than the conventional analyses that rely on fitting a common production technology for all firms regardless of their locations and thus confine cross-firm heterogeneity to differences in ω_{it} only.

Table 2.5 presents the decomposition results (across locations s) following (2.4.4). Because we just have three years of data, we perform the decomposition by pooling over the entire sample period. Thus, reported are the average decomposition results across 2002–2004. Also note that, for a fixed benchmark location κ, the decomposition is done for each s-location separately. For the benchmark/reference location κ, we choose the zipcode with the smallest mean production, i.e., $\kappa = \arg\min_s \overline{y}^s$, where $\overline{y}^s$ is defined as the time average of (2.4.2).[12] Therefore, the numbers (×100%) in Table 2.5 can be interpreted as the percentage differences between the chemicals manufacturers operating in various locations (s) versus those from the least-production-scale region (κ) in China. Because the reference location is fixed, the results are comparable across s.

Based on our estimates, the mean productivity differential is 1.866, which means that, compared to the location with the smallest scale of chemicals production, other locations are, on average, 187% more productive (or more effective in the input usage). The interquartile range of the average productivity differential spans from 1.652 and 2.086. Economically, these differences are large: firms that are located at the third quartile of the locational productivity distribution are about 43% more productive than firms at the first quartile. When we decompose the productivity differential into the technology and TFP differentials, on average, $\Delta\overline{\text{TECH}}^{s,\kappa}$ is 2.3 times as large as $\Delta\overline{\text{TFP}}^{s,\kappa}$ and accounts for about 69% of the total productivity differences across locations.[13] This suggests that the cross-location technological heterogeneity in China's chemicals industry explains most of the productivity differential and that the regional TFP differences are *relatively* more modest.

Table 2.5 also summarizes the locational productivity differential estimates from the standard location-invariant model. Given that this model assumes fixed coefficients (same technology for all firms), we cannot perform a decomposition here, and all cross-location variation in productivity is *a priori* attributed to TFP by design. Compared with our locationally-varying model, this model yields similar total productivity differentials across regions but, due to its inability to recognize technological differences, it grossly over-estimates cross-location differences in TFP.

To explore the spatial heterogeneity in the decomposition components, we plot the spatial distributions of $\Delta\overline{\text{TECH}}^{s,\kappa}$, $\Delta\overline{\text{TFP}}^{s,\kappa}$ and $\Delta\overline{\text{PROD}}^{s,\kappa}$ on the map in Figure 2.5. The spatial distribution of $\Delta\overline{\text{TECH}}^{s,\kappa}$ aligns remarkably with that of RTS in Figure 2.3. Noticeably, the regions of agglomeration in the chemicals industry (see Figure 2.1) tend to demonstrate large

[12] Obviously, the choice of a reference location is inconsequential because its role is effectively that of a normalization.

[13] That is, the ratio of $\Delta\overline{\text{TECH}}^{s,\kappa}$ to $\Delta\overline{\text{PROD}}^{s,\kappa}$ is 0.69.

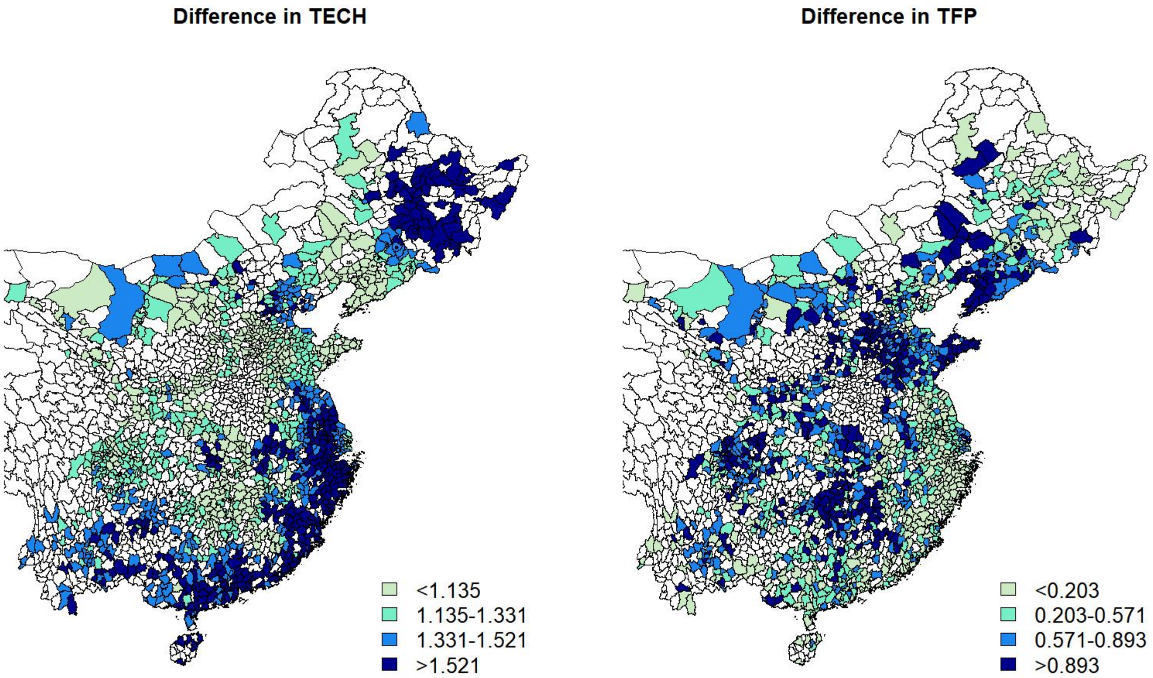

Figure 2.5. Locational Productivity Differential Decomposition Estimates Across Space
(Notes: The color shade cutoffs correspond to the first, second (median) and third quartiles)

technology differentials. In contrast, the spatial distribution of $\Delta\overline{\text{TFP}}^{s,\kappa}$ shows quite a different pattern, whereby the locations of large TFP (differentials) are less concentrated. Unlike with the $\Delta\overline{\text{TECH}}^{s,\kappa}$ map, the dark-shaded regions on the $\Delta\overline{\text{TFP}}^{s,\kappa}$ map are widely spread around and have no clear overlapping with the main agglomeration regions in the industry. The comparison between these two maps suggests that, at least for the Chinese chemicals manufacturing firms, the widely-documented agglomeration effects on firm productivity are associated more with the scale effects via production technology rather than the improvements in overall TFP. That is, by locating closer to other firms in the same industry, it may be easier for a firm to pick up production technologies and know-hows that improve productiveness of inputs technologically and thus expand the input requirement set corresponding to the firm's output level[14] *given* its total factor productivity. Instead, agglomeration effects that increase the effectiveness of transforming all factors into the outputs via available technology (by adopting better business practices or management strategies) may

[14] And more generally, shifting the *family* of firm's isoquants corresponding to a fixed level of ω_{it} toward the origin.

be less likely to spill among the Chinese manufacturers of chemicals. Importantly, if we *a priori* assume the fixed-coefficient production function common to all firms, the technological effects of agglomeration (via input elasticities) would be wrongly attributed to the TFP differentials.

2.6 Concluding Remarks

Although it is widely documented in the operations management literature that the firm's location matters for its performance, few empirical studies of operations efficiency explicitly control for it. This paper fills in this gap by providing a semiparametric methodology for the identification of production functions in which locational factors have heterogeneous effects on the firm's production technology and productivity evolution. Our approach is novel in that we explicitly model spatial variation in parameters in the production-function estimation. We generalize the popular Cobb-Douglas production function in a semiparametric fashion by writing the input elasticities and productivity parameters as unknown functions of the firm's geographic location. In doing so, not only do we render the production technology location-specific but also accommodate neighborhood influences on firm operations with the strength thereof depending on the distance between firms. Importantly, this enables us to examine the role of cross-location differences in explaining the variation in operational productivity among firms.

The proposed model is superior to the alternative SAR-type production-function formulations because it (i) explicitly estimates the locational variation in production functions, (ii) is readily reconcilable with the conventional production axioms and, more importantly, (iii) can be identified from the data by building on the popular proxy-variable methods, which we extend to incorporate locational heterogeneity in firm production. Our methodology provides a practical tool for examining the effects of agglomeration and technology spillovers on firm performance and will be most useful for empiricists focused on the analysis of operations efficiency/productivity and its "determinants."

Using the methods proposed in our paper, we can separate the effects of firm location on production technology from those on firm productivity and find evidence consistent with the conclusion that agglomeration economies affect the productivity of Chinese chemicals manufacturers mainly through the scale effects of production *technology* rather than the improvements in overall TFP. Comparing our flexible semiparametric model with the more conventional parametric model that postulates a common technology for all firms regardless of their location, we show that the latter does not provide an adequate representation of the industry and that the conclusion based on its results can be misleading. For managerial implications, our study re-emphasizes the importance of firm location for its operations efficiency in manufacturing industries. Our findings also suggest that hiring skilled labor has a larger productivity effect compared to other widely-discussed productivity-enhancing techniques, such as learning by exporting.

Chapter 3

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

3.1 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

[*] This chapter is based on Zhang and Malikov (2021).

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-Tobòn and Kumbhakar, 2015); others find economies of scale for medium and small banks as well (e.g., Malikov et al., 2015; Restrepo-Tobòn et al., 2015; Wheelock and Wilson, 2018).

With the sole exception (see below), 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.[1] 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. Since then, banks in the U.S. have 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). Thus, the portfolio of products offered by the modern banks

[1] 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).

is very different from that two decades ago, which underscores the importance of our paper.

Nontraditional 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). To 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.

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. For banks in the middle interquartile range of the cost distribution, as many as 82.3% exhibit positive economies of scope. For the top half of the distribution, the prevalence of significant scope economies is ≥97.5%. Even at the very bottom of cost distribution where the product diversification opportunities may not be as abundant or easily accessible, our test results suggest that roughly two thirds of banks (65%) enjoy scope-driven cost savings and those, who do not, largely exhibit scope invariance. We find *no* material evidence in support of scope *dis*economies. Our findings are in stark contrast with earlier studies.

The rest of the paper unfolds as follows. Section 3.2 discusses the theoretical framework. Section 3.3 describes our econometric model. Data are summarized in Section 3.4, following by Section 3.5 that reports the empirical results. We then conclude in Section 3.6.

3.2 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:

$$\mathscr{C}_t(\boldsymbol{Y},\boldsymbol{W},\boldsymbol{K}) = \min_{\boldsymbol{X}\geq\boldsymbol{0}}\left\{\boldsymbol{X}'\boldsymbol{W} \mid (\boldsymbol{X},\boldsymbol{K}) \text{ can produce } \boldsymbol{Y} \text{ at time } t\right\}, \tag{3.2.1}$$

where the arguments of cost function $\mathscr{C}_t(\cdot)$ are the output quantities $\boldsymbol{Y} \in \Re^M_+$, variable input prices $\boldsymbol{W} \in \Re^J_+$ and fixed input quantities $\boldsymbol{K} \in \Re^P_+$; and $\boldsymbol{X} \in \Re^J_+$ is the vector of variable input quantities. Importantly, the cost function in (3.2.1) 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 decreasing 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 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|notion of revenue diversification. Then, for some distribution weights $0 \leq \varpi_m^{\kappa} \leq 1$ such that $\sum_{\kappa} \varpi_m^{\kappa} = 1$ for all $m = 1, 2, 3$ and $\kappa \in \{A, B, C\}$, the bank is said to enjoy scope economies

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

at time t if

$$\sum_{\kappa \in \{A,B,C\}} \mathscr{C}_t\big(\varpi_1^\kappa Y_1, \varpi_2^\kappa Y_2, \varpi_3^\kappa Y_3\big) - \mathscr{C}_t\big(Y_1, Y_2, Y_3\big) > 0, \tag{3.2.2}$$

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 $\mathscr{C}_t(Y_1,0,0)+\mathscr{C}_t(0,Y_2,0)+\mathscr{C}_t(0,0,Y_3)-\mathscr{C}_t(Y_1,Y_2,Y_3) > 0$, 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 (3.2.2) with a pair of weights taking zero values for each counterfactual bank.

To further avoid excessive extrapolation, we restrict the choice of $\{\varpi_m\}$ 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 $\varpi_m^\kappa Y_m \geq \min\{Y_m\}$ for all $m = 1,2,3$ and $\kappa \in \{A,B,C\}$. The second constraint ensures that 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 $Y_{m'}$:

$$\min\left\{\frac{Y_m}{Y_{m'}}\right\} \leq \frac{\varpi_m^\kappa Y_m^* + \min\{Y_m\}}{\varpi_{m'}^\kappa Y_{m'}^* + \min\{Y_{m'}\}} \leq \max\left\{\frac{Y_m}{Y_{m'}}\right\}, \tag{3.2.3}$$

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 $\mathscr{S}_t$ (in proportions) is obtained by dividing the expression in (3.2.2) by $\mathscr{C}_t(Y_1, Y_2, Y_3)$:

$$\mathscr{S}_t = \frac{\sum_{\kappa\in\{A,B,C\}} \mathscr{C}_t\Big(\varpi_1^\kappa Y_1^* + \min\{Y_1\}, \varpi_2^\kappa Y_2^* + \min\{Y_2\}, \varpi_3^\kappa Y_3^* + \min\{Y_3\}\Big) - \mathscr{C}_t\big(Y_1, Y_2, Y_3\big)}{\mathscr{C}_t\big(Y_1, Y_2, Y_3\big)}, \tag{3.2.4}$$

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 $\mathscr{S}_t$ provide evidence of scope economies (*dis*economies); while a zero value suggests scope invariance of the bank's cost structure.

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

$$\mathscr{S}_t^* = \min_{\{\varpi_m^\kappa\}} \mathscr{S}_t\big(\varpi_m^\kappa;\ m = 1,2,3; \kappa \in \{A,B,C\}\big)\,. \tag{3.2.5}$$

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 output space in a given year. Thus, the main hypothesis of interest is as follows.

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

3.3 Empirical Model

We estimate the bank's dual variable cost function $\mathscr{C}_t(\cdot)$ 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 $\boldsymbol{V}_{it} = (\boldsymbol{Y}_{it}', \boldsymbol{W}_{it}', \boldsymbol{K}_{it}')'$ be the vector of (strictly exogenous) cost-function regressors. We use lower case of C_{it} and $\boldsymbol{V}_{it}$ in the following to denote the log transformations of the variables: e.g., $\boldsymbol{v}_{it} = \ln \boldsymbol{V}_{it}$. Letting the bank's variable cost structure be of the translog[3] form and described by a location-scale model à la Koenker and Bassett (1982) extended to accommodate bank

[3] Quadratic log-polynomial.

fixed effects and time-varying coefficients, we have

$$c_{it} = \left[\beta_0 + \beta_0^* L(t)\right] + \left[\boldsymbol{\beta}_1 + \boldsymbol{\beta}_1^* L(t)\right]' \boldsymbol{v}_{it} + \tfrac{1}{2}\left[\boldsymbol{\beta}_2 + \boldsymbol{\beta}_2^* L(t)\right]' \mathrm{vec}\left(\boldsymbol{v}_{it}\boldsymbol{v}_{it}'\right) + \lambda_i + u_{it}, \tag{3.3.1}$$

with

$$u_{it} = \Big(\left[\gamma_0 + \gamma_0^* S(t)\right] + \left[\boldsymbol{\gamma}_1 + \boldsymbol{\gamma}_1^* S(t)\right]' \boldsymbol{v}_{it} + \tfrac{1}{2}\left[\boldsymbol{\gamma}_2 + \boldsymbol{\gamma}_2^* S(t)\right]' \mathrm{vec}\left(\boldsymbol{v}_{it}\boldsymbol{v}_{it}'\right) + \sigma_i\Big)\varepsilon_{it}, \tag{3.3.2}$$

where $\left(\beta_0, \boldsymbol{\beta}_1', \boldsymbol{\beta}_2', \beta_0^*, \boldsymbol{\beta}_1^{*\prime}, \boldsymbol{\beta}_2^{*\prime}\right)'$ are unknown location-function coefficients; $\left(\gamma_0, \boldsymbol{\gamma}_1', \boldsymbol{\gamma}_2', \gamma_0^*, \boldsymbol{\gamma}_1^{*\prime}, \boldsymbol{\gamma}_2^{*\prime}\right)'$ 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 $\boldsymbol{\beta}_1 + \boldsymbol{\beta}_1^* L(t)$ and quadratic slopes $\boldsymbol{\beta}_2 + \boldsymbol{\beta}_2^* L(t)$, 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 (3.3.1)–(3.3.2) is a generalization of the popular translog cost-funct-ion 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*) $\varepsilon_{it} \perp \boldsymbol{v}_{it}$ with the normalizations that

$\mathbb{E}[\varepsilon_{it}] = 0$ and $\mathbb{E}[|\varepsilon_{it}|] = 1$; and (*iii*) $\Pr\Big[[\gamma_0 + \gamma_0^* S(t)] + [\boldsymbol{\gamma}_1 + \boldsymbol{\gamma}_1^* S(t)]'\boldsymbol{v}_{it} + \frac{1}{2}[\boldsymbol{\gamma}_2 + \boldsymbol{\gamma}_2^* S(t)]' \times \mathrm{vec}\big(\boldsymbol{v}_{it}\boldsymbol{v}_{it}'\big) + \sigma_i > 0\Big] = 1$. Then, for any given quantile index $\tau \in (0,1)$, the τth conditional quantile function of the log-cost c_{it} implied by (3.3.1)–(3.3.2) is

$$\begin{aligned}\mathcal{Q}_c[\tau|\boldsymbol{v}_{it}] = {} & \overbrace{\Big[\beta_0 + \gamma_0 q_\tau + \beta_0^* L(t) + \gamma_0^* S(t) q_\tau\Big]}^{t\text{-varying quantile intercept}} + \overbrace{\Big[\boldsymbol{\beta}_1 + \boldsymbol{\gamma}_1 q_\tau + \boldsymbol{\beta}_1^* L(t) + \boldsymbol{\gamma}_1^* S(t) q_\tau\Big]'}^{t\text{-varying linear quantile slopes}} \boldsymbol{v}_{it} + {} \\ & \underbrace{\frac{1}{2}\Big[\boldsymbol{\beta}_2 + \boldsymbol{\gamma}_2 q_\tau + \boldsymbol{\beta}_2^* L(t) + \boldsymbol{\gamma}_2^* S(t) q_\tau\Big]'}_{t\text{-varying quadratic quantile slopes}} \mathrm{vec}\big(\boldsymbol{v}_{it}\boldsymbol{v}_{it}'\big) + \underbrace{\Big[\lambda_i + \sigma_i q_\tau\Big]}_{\text{individual quantile fixed effect}}\end{aligned} \tag{3.3.3}$$

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

The translog cost model in (3.3.3) 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 (3.3.3) can be rewritten compactly as

$$\mathcal{Q}_c[\tau|\boldsymbol{v}_{it}] \equiv \alpha_0(\tau, t) + \boldsymbol{\alpha}_{\mathbf{1}}(\tau, t)'\boldsymbol{v}_{it} + \frac{1}{2}\boldsymbol{\alpha}_{\mathbf{2}}(\tau, t)'\mathrm{vec}\big(\boldsymbol{v}_{it}\boldsymbol{v}_{it}'\big) + \mu_{i,\tau}, \tag{3.3.4}$$

with the “alpha” coefficients corresponding to the bracketed expressions in (3.3.3) 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 (3.3.3) as opposed to postulating a quantile regression à la (3.3.4) *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 (3.3.4), 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 (3.3.1)–(3.3.2). 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 $\kappa = 1, \ldots, T$, define the dummy variable $D_{\kappa,t}$ that is equal to 1 in the κth 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, $\boldsymbol{\beta}_1$, γ_0 and $\boldsymbol{\gamma}_1$ are naturally interpretable as "ref-

[4] For instance, in our empirical application $n > 7{,}500$.

erence" coefficients in time period $t = 1$. Then, a feasible analogue of the τth conditional cost quantile in (3.3.3) is given by

$$Q_c\left[\tau|\boldsymbol{v}_{it}\right] = \left[\beta_0 + \gamma_0 q_\tau + \sum_\kappa (\eta_\kappa + \theta_\kappa q_\tau) D_{\kappa,t}\right] + \left[\boldsymbol{\beta}_1 + \boldsymbol{\gamma}_1 q_\tau + \sum_\kappa (\boldsymbol{\beta}_1^* \eta_\kappa + \boldsymbol{\gamma}_1^* \theta_\kappa q_\tau) D_{\kappa,t}\right]' \boldsymbol{v}_{it} + \\ \frac{1}{2}\left[\boldsymbol{\beta}_2 + \boldsymbol{\gamma}_2 q_\tau + \sum_\kappa (\boldsymbol{\beta}_2^* \eta_\kappa + \boldsymbol{\gamma}_2^* \theta_\kappa q_\tau) D_{\kappa,t}\right]' \text{vec}\left(\boldsymbol{v}_{it} \boldsymbol{v}_{it}'\right) + \left[\lambda_i + \sigma_i q_\tau\right]. \tag{3.3.5}$$

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 using a series of time dummies, we obtain the time-varying slope coefficients on $\boldsymbol{v}_{it}$ 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 $\boldsymbol{v}_{it}$. 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. This is important because, by avoiding a fully saturated specification that is equivalent to sample-splitting into cross-sections, we are able to accommodate time-invariant individual fixed effects in the model, the estimation of which requires the cross-time "within" variation.[5]

Although the estimation of (3.3.3) 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 (3.3.1)–(3.3.2) 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 (3.3.3). In what follows, we describe this pro-

[5] Incidentally, the nonlinearity of our model is also the reason why we do not use the Canay (2011) fixed-effects quantile regression estimator which provides an alternative to Machado and Santos Silva (2019) with the difference being that the former uses the check-function-based estimator as opposed to moment-based. Given that the L_1 optimization used in the check-function-based estimations is not as adept to handling nonlinearities in parameters and, more importantly, to the presence of many dummy variables, we opt for the L_2 moment-based estimator.

cedure in detail.

3.3.1 Estimation Procedure

First, for ease of notation, we define $\boldsymbol{D}_t = [D_{2,t},\ldots,D_{T,t}]'$, $\boldsymbol{\eta} = [\eta_2,\ldots,\eta_T]'$ and $\boldsymbol{\theta} = [\theta_2,\ldots,\theta_T]'$.

Step 1. We first estimate parameters of the location function. Under the assumption (*ii*), from (3.3.1) it follows that the conditional mean function of the log-cost c_{it} is

$$\mathbb{E}\left[c_{it}|\boldsymbol{v}_{it},\boldsymbol{D}_t\right] = \beta_0+\sum_{\kappa}\eta_\kappa D_{\kappa,t}+\left[\boldsymbol{\beta}_1+\boldsymbol{\beta}_1^*\sum_{\kappa}\eta_\kappa D_{\kappa,t}\right]'\boldsymbol{v}_{it}+\frac{1}{2}\left[\boldsymbol{\beta}_2+\boldsymbol{\beta}_2^*\sum_{\kappa}\eta_\kappa D_{\kappa,t}\right]'\text{vec}\left(\boldsymbol{v}_{it}\boldsymbol{v}_{it}'\right)+\lambda_i, \tag{3.3.6}$$

which can be consistently estimated in the within-transformed form via nonlinear least squares after purging additive location fixed effects.[6]

Given the nonlinearity and high dimensionality of (3.3.6) in parameters, we estimate the slope coefficients via concentration by noticing that, conditional on $\boldsymbol{\eta}$, this mean regression is linear in $[\boldsymbol{\beta}_1',\boldsymbol{\beta}_2',\boldsymbol{\beta}_1^{*\prime},\boldsymbol{\beta}_2^{*\prime}]'$ 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

$$\begin{aligned} M(\boldsymbol{\eta}) = \sum_i\sum_t \Big[& c_{it}-\overline{c}_i-\boldsymbol{\eta}'\left(\boldsymbol{D}_t-\overline{\boldsymbol{D}}\right)-\left(\boldsymbol{v}_{it}-\overline{\boldsymbol{v}}_i\right)'\boldsymbol{\beta}_1\left(\boldsymbol{\eta}\right)\; - \\ & \left(\boldsymbol{\eta}'\boldsymbol{D}_t\cdot\boldsymbol{v}_{it}-\overline{\boldsymbol{\eta}'\boldsymbol{D}_t\cdot\boldsymbol{v}_i}\right)'\boldsymbol{\beta}_1^*\left(\boldsymbol{\eta}\right)-\frac{1}{2}\left(\text{vec}\left(\boldsymbol{v}_{it}\boldsymbol{v}_{it}'\right)-\overline{\text{vec}\left(\boldsymbol{v}_i\boldsymbol{v}_i'\right)}\right)\boldsymbol{\beta}_2\left(\boldsymbol{\eta}\right)\; - \\ & \frac{1}{2}\left(\boldsymbol{\eta}'\boldsymbol{D}_t\cdot\text{vec}\left(\boldsymbol{v}_{it}\boldsymbol{v}_{it}'\right)-\overline{\boldsymbol{\eta}'\boldsymbol{D}_t\cdot\text{vec}\left(\boldsymbol{v}_i\boldsymbol{v}_i'\right)}\right)'\boldsymbol{\beta}_2^*\left(\boldsymbol{\eta}\right)\Big]^2, \end{aligned} \tag{3.3.7}$$

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

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

[6] Note that, although (3.3.6) 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.

in the location functions are

$$\widehat{\boldsymbol{\eta}} = \arg\min_{\boldsymbol{\eta}} M(\boldsymbol{\eta}) \quad \text{and} \quad \widehat{\boldsymbol{\beta}}_1 = \boldsymbol{\beta}_1\left(\widehat{\boldsymbol{\eta}}\right), \quad \widehat{\boldsymbol{\beta}}_2 = \boldsymbol{\beta}_2\left(\widehat{\boldsymbol{\eta}}\right), \quad \widehat{\boldsymbol{\beta}}_1^* = \boldsymbol{\beta}_1^*\left(\widehat{\boldsymbol{\eta}}\right), \quad \widehat{\boldsymbol{\beta}}_2^* = \boldsymbol{\beta}_2^*\left(\widehat{\boldsymbol{\eta}}\right). \tag{3.3.8}$$

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

$$\widehat{\beta}_0 = \frac{1}{nT}\sum_i\sum_t \left(c_{it} - \widehat{\boldsymbol{\eta}}'\boldsymbol{D}_t - \left[\widehat{\boldsymbol{\beta}}_1 + \widehat{\boldsymbol{\eta}}'\boldsymbol{D}_t \cdot \widehat{\boldsymbol{\beta}}_1^*\right]'\boldsymbol{v}_{it} - \frac{1}{2}\left[\widehat{\boldsymbol{\beta}}_2 + \widehat{\boldsymbol{\eta}}'\boldsymbol{D}_t \cdot \widehat{\boldsymbol{\beta}}_2^*\right]'\operatorname{vec}\left(\boldsymbol{v}_{it}\boldsymbol{v}_{it}'\right)\right), \tag{3.3.9}$$

$$\widehat{\lambda}_i = \frac{1}{T}\sum_t \left(c_{it} - \widehat{\beta}_0 - \widehat{\boldsymbol{\eta}}'\boldsymbol{D}_t - \left[\widehat{\boldsymbol{\beta}}_1 + \widehat{\boldsymbol{\eta}}'\boldsymbol{D}_t \cdot \widehat{\boldsymbol{\beta}}_1^*\right]'\boldsymbol{v}_{it} - \frac{1}{2}\left[\widehat{\boldsymbol{\beta}}_2 + \widehat{\boldsymbol{\eta}}'\boldsymbol{D}_t \cdot \widehat{\boldsymbol{\beta}}_2^*\right]'\operatorname{vec}\left(\boldsymbol{v}_{it}\boldsymbol{v}_{it}'\right)\right) \; \forall i. \tag{3.3.10}$$

Hence, the residual estimator is

$$\widehat{u}_{it} = c_{it} - \widehat{\beta}_0 - \widehat{\boldsymbol{\eta}}'\boldsymbol{D}_t - \left[\widehat{\boldsymbol{\beta}}_1 + \widehat{\boldsymbol{\eta}}'\boldsymbol{D}_t \cdot \widehat{\boldsymbol{\beta}}_1^*\right]'\boldsymbol{v}_{it} - \frac{1}{2}\left[\widehat{\boldsymbol{\beta}}_2 + \widehat{\boldsymbol{\eta}}'\boldsymbol{D}_t \cdot \widehat{\boldsymbol{\beta}}_2^*\right]'\operatorname{vec}\left(\boldsymbol{v}_{it}\boldsymbol{v}_{it}'\right) - \widehat{\lambda}_i. \tag{3.3.11}$$

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

$$\mathbb{E}\left[|u_{it}||\boldsymbol{v}_{it}, \boldsymbol{D}_t\right] = \gamma_0 + \sum_\kappa \theta_\kappa D_{\kappa,t} + \left[\boldsymbol{\gamma}_1 + \boldsymbol{\gamma}_1^* \sum_\kappa \theta_\kappa D_{\kappa,t}\right]'\boldsymbol{v}_{it} + \frac{1}{2}\left[\boldsymbol{\gamma}_2 + \boldsymbol{\gamma}_2^* \sum_\kappa \theta_\kappa D_{\kappa,t}\right]'\operatorname{vec}\left(\boldsymbol{v}_{it}\boldsymbol{v}_{it}'\right) + \sigma_i, \tag{3.3.12}$$

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

$$\begin{aligned} M(\boldsymbol{\theta}) = \sum_i\sum_t \Bigg[& |\widehat{u}_{it}| - |\overline{\widehat{u}}_i| - \boldsymbol{\theta}'\left(\boldsymbol{D}_t - \overline{\boldsymbol{D}}\right) - \left(\boldsymbol{v}_{it} - \overline{\boldsymbol{v}}_i\right)'\boldsymbol{\gamma}_1(\boldsymbol{\theta}) \; - \\ & \left(\boldsymbol{\theta}'\boldsymbol{D}_t \cdot \boldsymbol{v}_{it} - \overline{\boldsymbol{\theta}'\boldsymbol{D}_t \cdot \boldsymbol{v}_i}\right)'\boldsymbol{\gamma}_1^*(\boldsymbol{\theta}) - \frac{1}{2}\left(\operatorname{vec}\left(\boldsymbol{v}_{it}\boldsymbol{v}_{it}'\right) - \overline{\operatorname{vec}\left(\boldsymbol{v}_i\boldsymbol{v}_i'\right)}\right)\boldsymbol{\gamma}_2(\boldsymbol{\theta}) \; - \\ & \frac{1}{2}\left(\boldsymbol{\theta}'\boldsymbol{D}_t \cdot \operatorname{vec}\left(\boldsymbol{v}_{it}\boldsymbol{v}_{it}'\right) - \overline{\boldsymbol{\theta}'\boldsymbol{D}_t \cdot \operatorname{vec}\left(\boldsymbol{v}_i\boldsymbol{v}_i'\right)}\right)'\boldsymbol{\gamma}_2^*(\boldsymbol{\theta}) \Bigg]^2, \end{aligned} \tag{3.3.13}$$

and the corresponding profiled estimators given by $[\boldsymbol{\gamma}_1(\boldsymbol{\theta})', \boldsymbol{\gamma}_2(\boldsymbol{\theta})', \boldsymbol{\gamma}_1^*(\boldsymbol{\theta})', \boldsymbol{\gamma}_2^*(\boldsymbol{\theta})']' = \left(\sum_i \sum_t \mathscr{X}_{it}\mathscr{X}_{it}'\right)^{-1} \times \sum_i \sum_t \mathscr{X}_{it}\mathscr{Y}_{it}^{\dagger}$, where $\mathscr{X}_{it} = \left[\left(\boldsymbol{v}_{it} - \overline{\boldsymbol{v}}_i\right)', \left(\boldsymbol{\theta}'\boldsymbol{D}_t \cdot \boldsymbol{v}_{it} - \overline{\boldsymbol{\theta}'\boldsymbol{D}_t \cdot \boldsymbol{v}_i}\right)', \frac{1}{2}\left(\mathrm{vec}\left(\boldsymbol{v}_{it}\boldsymbol{v}_{it}'\right) - \overline{\mathrm{vec}\left(\boldsymbol{v}_i\boldsymbol{v}_i'\right)}\right), \frac{1}{2}\left(\boldsymbol{\theta}'\boldsymbol{D}_t \cdot \mathrm{vec}\left(\boldsymbol{v}_{it}\boldsymbol{v}_{it}'\right) - \overline{\boldsymbol{\theta}'\boldsymbol{D}_t \cdot \mathrm{vec}\left(\boldsymbol{v}_i\boldsymbol{v}_i'\right)}\right)'\right.$ and $\mathscr{Y}_{it}^{\dagger} = |\widehat{u}_{it}| - |\overline{\widehat{u}}_i| - \boldsymbol{\theta}'\left(\boldsymbol{D}_t - \overline{\boldsymbol{D}}\right)$, the nonlinear fixed-effects estimators of the scale-function slope coefficients $[\boldsymbol{\theta}', \boldsymbol{\gamma}_1', \boldsymbol{\gamma}_2', \boldsymbol{\gamma}_3', \boldsymbol{\gamma}_4']'$ are

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} M(\boldsymbol{\theta}) \quad \text{and} \quad \widehat{\boldsymbol{\gamma}}_1 = \boldsymbol{\gamma}_1\left(\widehat{\boldsymbol{\theta}}\right), \quad \widehat{\boldsymbol{\gamma}}_2 = \boldsymbol{\gamma}_2\left(\widehat{\boldsymbol{\theta}}\right), \quad \widehat{\boldsymbol{\gamma}}_1^* = \boldsymbol{\gamma}_1^*\left(\widehat{\boldsymbol{\theta}}\right), \quad \widehat{\boldsymbol{\gamma}}_2^* = \boldsymbol{\gamma}_2^*\left(\widehat{\boldsymbol{\theta}}\right). \quad (3.3.14)$$

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

$$\widehat{\gamma}_0 = \frac{1}{nT}\sum_i \sum_t \left(|\widehat{u}_{it}| - \widehat{\boldsymbol{\theta}}'\boldsymbol{D}_t - \left[\widehat{\boldsymbol{\gamma}}_1 + \widehat{\boldsymbol{\theta}}'\boldsymbol{D}_t \cdot \widehat{\boldsymbol{\gamma}}_1^*\right]'\boldsymbol{v}_{it} - \frac{1}{2}\left[\widehat{\boldsymbol{\gamma}}_2 + \widehat{\boldsymbol{\theta}}'\boldsymbol{D}_t \cdot \widehat{\boldsymbol{\gamma}}_2^*\right]'\mathrm{vec}\left(\boldsymbol{v}_{it}\boldsymbol{v}_{it}'\right)\right), \quad (3.3.15)$$

$$\widehat{\sigma}_i = \frac{1}{T}\sum_t \left(|\widehat{u}_{it}| - \widehat{\gamma}_0 - \widehat{\boldsymbol{\theta}}'\boldsymbol{D}_t - \left[\widehat{\boldsymbol{\gamma}}_1 + \widehat{\boldsymbol{\theta}}'\boldsymbol{D}_t \cdot \widehat{\boldsymbol{\gamma}}_1^*\right]'\boldsymbol{v}_{it} - \frac{1}{2}\left[\widehat{\boldsymbol{\gamma}}_2 + \widehat{\boldsymbol{\theta}}'\boldsymbol{D}_t \cdot \widehat{\boldsymbol{\gamma}}_2^*\right]'\mathrm{vec}\left(\boldsymbol{v}_{it}\boldsymbol{v}_{it}'\right)\right) \forall i. \quad (3.3.16)$$

Step 3. For any given quantile index $0 < \tau < 1$ of interest, we next estimate the unconditional quantile of ε_{it}. From (3.3.2), we have the conditional quantile function of u_{it}:

$$\mathscr{Q}_u\left[\tau|\boldsymbol{v}_{it}, \boldsymbol{D}_t\right] = \left(\gamma_0 + \sum_\kappa \theta_\kappa D_{\kappa,t} + \left[\boldsymbol{\gamma}_1 + \boldsymbol{\gamma}_1^* \sum_\kappa \theta_\kappa D_{\kappa,t}\right]'\boldsymbol{v}_{it} + \frac{1}{2}\left[\boldsymbol{\gamma}_2 + \boldsymbol{\gamma}_2^* \sum_\kappa \theta_\kappa D_{\kappa,t}\right]'\mathrm{vec}\left(\boldsymbol{v}_{it}\boldsymbol{v}_{it}'\right) + \sigma_i\right) q_\tau. \quad (3.3.17)$$

We therefore can estimate q_τ via a univariate quantile regression (with no intercept) via

$$\widehat{q}_\tau = \arg\min_q \sum_i \sum_t \rho_\tau\left\{\widehat{u}_{it} - \left(\widehat{\gamma}_0 + \widehat{\boldsymbol{\theta}}'\boldsymbol{D}_t + \left[\widehat{\boldsymbol{\gamma}}_1 + \widehat{\boldsymbol{\theta}}'\boldsymbol{D}_t \cdot \widehat{\boldsymbol{\gamma}}_1^*\right]'\boldsymbol{v}_{it} + \frac{1}{2}\left[\widehat{\boldsymbol{\gamma}}_2 + \widehat{\boldsymbol{\theta}}'\boldsymbol{D}_t \cdot \widehat{\boldsymbol{\gamma}}_2^*\right]'\mathrm{vec}\left(\boldsymbol{v}_{it}\boldsymbol{v}_{it}'\right) + \widehat{\sigma}_i\right) q\right\}, \quad (3.3.18)$$

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

With all unknown parameters now estimated, we can construct the estimator of the feasible analogue of the τth conditional quantile of the log-cost in (3.3.3).

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.2.

3.4 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 discussed in the introduction, our primary focus is on the post-financial-crisis period.

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 58,021 observations for 7,583 banks. We deflate all nominal variables to the 2005 U.S. dollars using the consumer price index. Consistent with the widely used intermediation approach of Sealey and Lindley (1977), we define the variables as follows.

The two traditional 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). We also include the output measure of nontraditional bank operations such trading and investment services, Y_3. Since non-interest income is heavily influenced by off-balance-sheet activities (Clark and Siems, 2002), we follow the literature in using the non-interest income minus service charges on deposits as a measure of off-balance-sheet income (e.g., Wheelock and Wilson, 2012, 2018; Malikov et al., 2015).

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,

Table 3.1. Data Summary Statistics

Variables	Mean	1st Qu.	Median	3rd Qu.
C	39,692.94	1,874.71	3,911.11	8,946.57
Y_1	1,140,365.00	43,481.64	97,922.83	236,489.00
Y_2	417,858.60	11,492.15	29,088.64	72,806.76
Y_3	28,890.97	145.65	480.42	1,697.59
W_1	61.06	15.27	22.50	38.03
W_2	58.33	47.02	54.67	65.56
W_3	0.740	0.362	0.589	0.976
K_1	229,008.90	8,090.69	16,506.67	37,521.77
K_2	0.026	0.008	0.016	0.032

C – total variable costs; Y_1 – total loans; Y_2 – total securities; Y_3 – off-balance-sheet output; 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. 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 and K_2 are unit-free.

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 (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 a proxy measure of its credit risk. Specifically, we use the ratio of nonperforming assets to total assets K2, which reflects the quality of assets held by the bank.[7] Controlling for nonperforming outputs when modeling bank costs is imperative because lower-quality assets generally require more resources to manage a higher-level (e.g., see Hughes and Mester, 2013). Following the literature, we define nonperforming assets as a sum of (*i*) total loans and lease financing receivables past due 30 days or more and still accruing, (*ii*) total loans and lease financing receivables not accruing, (*iii*) other real estate owned and (*iv*) charge-offs on past-due loans and leases. The loss provision is measured using the total provision for loan and lease losses. Table 3.1 provides summary statistics for these variables.

3.5 Empirical Results

This section reports the results based on our time-varying-coefficient fixed-effects quantile model of bank cost in (3.3.1)–(3.3.2) 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 tech-

[7] While we denote this variable as a "K," we do not conceptualize it as a quasi-fixed input quantity analogous to K1.

nological progress. The summary of usual cost elasticities are reported in the Appendix.

3.5.1 Scope Economies

As discussed in Section 3.2, 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 (3.2.4)–(3.2.5), we replace $\mathscr{C}_t(\cdot)$ with the exponentiated quantile function of the log-cost $\mathscr{Q}_c(\tau|\cdot)$ 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

$$\mathscr{S}_t(\tau) = \frac{\sum_{\kappa} \exp\left[\mathscr{Q}_c\Big(\tau|\varpi_1^{\kappa} Y_1^* + \min\{Y_1\}, \varpi_2^{\kappa} Y_2^* + \min\{Y_2\}, \varpi_3^{\kappa} Y_3^* + \min\{Y_3\}, t\Big)\right] - \exp\left[\mathscr{Q}_c\big(\tau|Y_1, Y_2, Y_3, t\big)\right]}{\exp\left[\mathscr{Q}_c\big(\tau|Y_1, Y_2, Y_3, t\big)\right]}. \tag{3.5.1}$$

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 $\mathscr{S}_t(\tau)$, 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 $c = f(\boldsymbol{v}) + \epsilon$ with $\mathbb{E}[\epsilon|\boldsymbol{v}] = 0$, and recall that upper/lower-case variables are in levels/logs. It then trivially follows that $\mathbb{E}[C|\boldsymbol{v}] = \exp\{f(\boldsymbol{v})\}\mathbb{E}[\exp\{\epsilon\}|\boldsymbol{v}]$ which generally diverges from $\exp\{f(\boldsymbol{v})\}$ by a multiplicative function of $\boldsymbol{v}$. Since cost counterfactuals in $\mathscr{S}_t(\tau)$ admit different “$\boldsymbol{v}$” 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 $\boldsymbol{v}$ 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

Table 3.2. Cost Subadditivity Estimates

Cost	*Point Estimates*				*Inference Categories, %*			
Quantiles (τ)	Mean	1st Qu.	Median	3rd Qu.	$= \mathbf{0}$	$\neq 0$	$> \mathbf{0}$	≤ 0
$\mathcal{Q}(0.10)$	0.252	0.144	0.239	0.334	**37.97**	62.03	**64.96**	35.04
	(0.09, 0.375)	(–0.025, 0.293)	(0.096, 0.34)	(0.185, 0.428)				
$\mathcal{Q}(0.25)$	0.297	0.202	0.282	0.367	**20.27**	79.73	**82.27**	17.73
	(0.144, 0.403)	(0.063, 0.315)	(0.144, 0.377)	(0.214, 0.467)				
$\mathcal{Q}(0.50)$	0.361	0.280	0.346	0.421	**3.25**	96.75	**97.52**	2.48
	(0.241, 0.482)	(0.179, 0.4)	(0.229, 0.468)	(0.286, 0.558)				
$\mathcal{Q}(0.75)$	0.421	0.343	0.409	0.482	**0.98**	99.02	**99.22**	0.78
	(0.303, 0.587)	(0.25, 0.499)	(0.296, 0.587)	(0.33, 0.664)				
$\mathcal{Q}(0.90)$	0.457	0.379	0.447	0.520	**0.79**	99.21	**99.38**	0.62
	(0.334, 0.644)	(0.273, 0.528)	(0.325, 0.624)	(0.369, 0.734)				

The left panel summarizes point estimates of $\mathscr{S}_t^*(\tau)$ with the corresponding two-sided 95% bias-corrected confidence intervals in parentheses. Each bank-year is classified as exhibiting scope economies [$\mathscr{S}_t^*(\tau) > 0$] vs. non-economies [$\mathscr{S}_t^*(\tau) \leq 0$] and scope invariance [$\mathscr{S}_t^*(\tau) = 0$] vs. scope non-invariance [$\mathscr{S}_t^*(\tau) \neq 0$] 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.

the equivariance of quantiles to monotone transformations, viz. $\mathcal{Q}_C[\tau|\boldsymbol{v}] = \mathcal{Q}_{\exp\{c\}}[\tau|\boldsymbol{v}] = \exp\{\mathcal{Q}_c[\tau|\boldsymbol{v}]\}$ (e.g., see Koenker, 2005).

Now, recall that $\mathscr{S}_t(\tau)$ depends on the choice of $\{\varpi_m^\kappa\}$, which we circumvent by choosing weights that yield the smallest cost subadditivity measure for a given cost quantile τ in the admissible region: $\mathscr{S}_t^*(\tau)$. 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 3.2 summarizes such point estimates of $\mathscr{S}_t^*(\tau)$ 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 $\mathscr{S}_t^*(\tau)$ estimates corresponding to a given τ.)

The two hypotheses of particular interest here are (*i*) $\mathbb{H}_0 : \mathscr{S}_t^*(\tau) \leq 0$ v. $\mathbb{H}_1 : \mathscr{S}_t^*(\tau) > 0$ and (*ii*) $\mathbb{H}_0 : \mathscr{S}_t^*(\tau) = 0$ v. $\mathbb{H}_1 : \mathscr{S}_t^*(\tau) \neq 0$. Both tests are essentially the same, except for the one- or two-sided alternatives. Although the (i,t) index on outputs is suppressed in (3.5.1), 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 3.2 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 [$\mathscr{S}_t^*(\tau) > 0$] vs. scope non-economies [$\mathscr{S}_t^*(\tau) \leq 0$] and the banks whose cost structure that exhibits scope invariance [$\mathscr{S}_t^*(\tau) = 0$] vs. scope non-invariance [$\mathscr{S}_t^*(\tau) \neq 0$].

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, as many as 82.3% exhibit positive economies of scope. For the top half of the distribution (median or higher), the prevalence of significant scope economies is ≥97.5%. 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 two thirds of banks (65%) enjoy scope-driven cost savings and those, who do not, largely exhibit scope invariance. Figure 3.1 provides a graphic illustration of these results. For each considered cost quantile τ, the figure shows a scatterplot of the $\mathscr{S}_t^*(\tau)$ 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 3.1, it is evident that the share of banks enjoying significant scope economies is growing with the quantile of conditional cost distribution.

Overall, having accounted for three-way heterogeneity across banks in a pursuit of robust estimates of bank cost subadditivity, we find *no* notable empirical evidence in support of scope *dis*economies.[8] 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 and Wilson, 2001). Besides our reliance on the more robust estimation methodology, the qualitative differences between our and prior findings

[8] In fact, formally testing if $\mathscr{S}_t^*(\tau) < 0$ reveals that only at most 0.02% of banks show significantly negative scope economies, and this only pertains to the bottom 0.10th cost quantile.

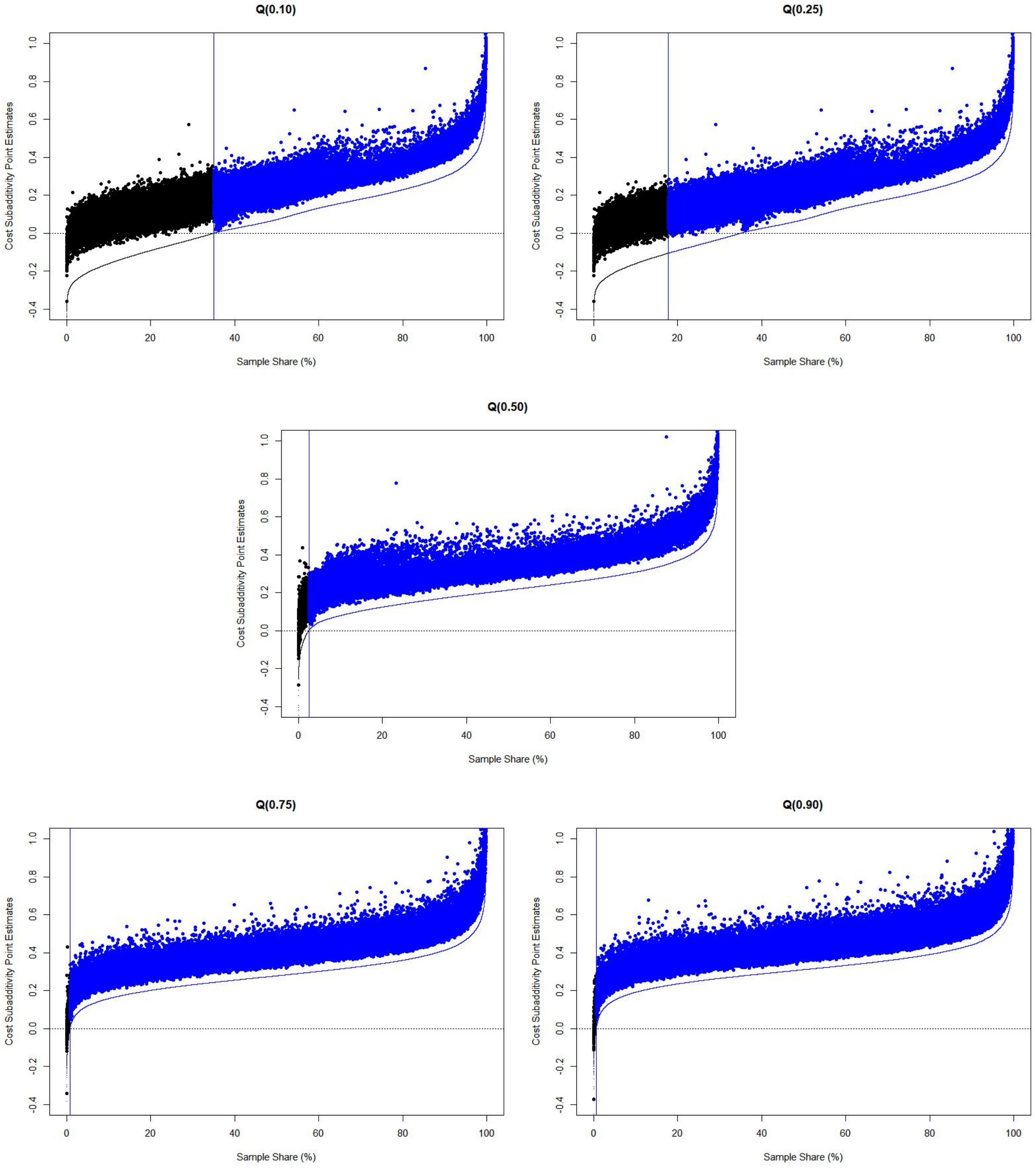

Figure 3.1. The One-Sided 95% Lower Bounds (solid lines) of the Cost Subadditivity Point Estimates (scatter points)

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 involving derivatives, securitization and mortgages as well as as other technological cost-saving advancements.

Although, the subadditivity measure does not directly quantify the *magnitude* of scope econ-omies 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 $\mathscr{S}_t^*(\tau)$ 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|notion 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 3.2, the mean estimates of cost subadditivity ranges from 0.252 to 0.457 depending on the conditional cost quantile. This suggests, on average, the potential for a 25–46% 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 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 box-plots and kernel density plots of the $\mathscr{S}_t^*(\tau)$ estimates, respectively graphed in Figure 3.2 and Figure 3.3. These enable us to compare distributions of the cost subadditivity estimates as opposed to merely focusing on marginal moments. Consistent with our earlier discussion, both figures indicate that large-scale banks lying in the upper quantiles of the cost distribution appear to enjoy 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.

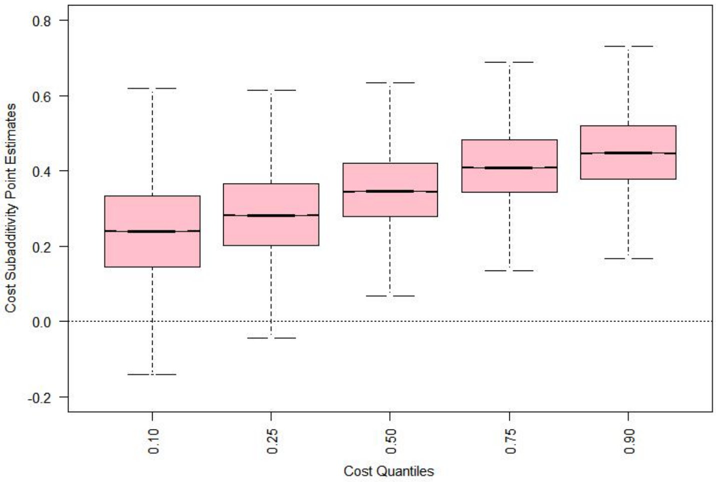

Figure 3.2. Cost Subadditivity Estimates Across Cost Quantiles

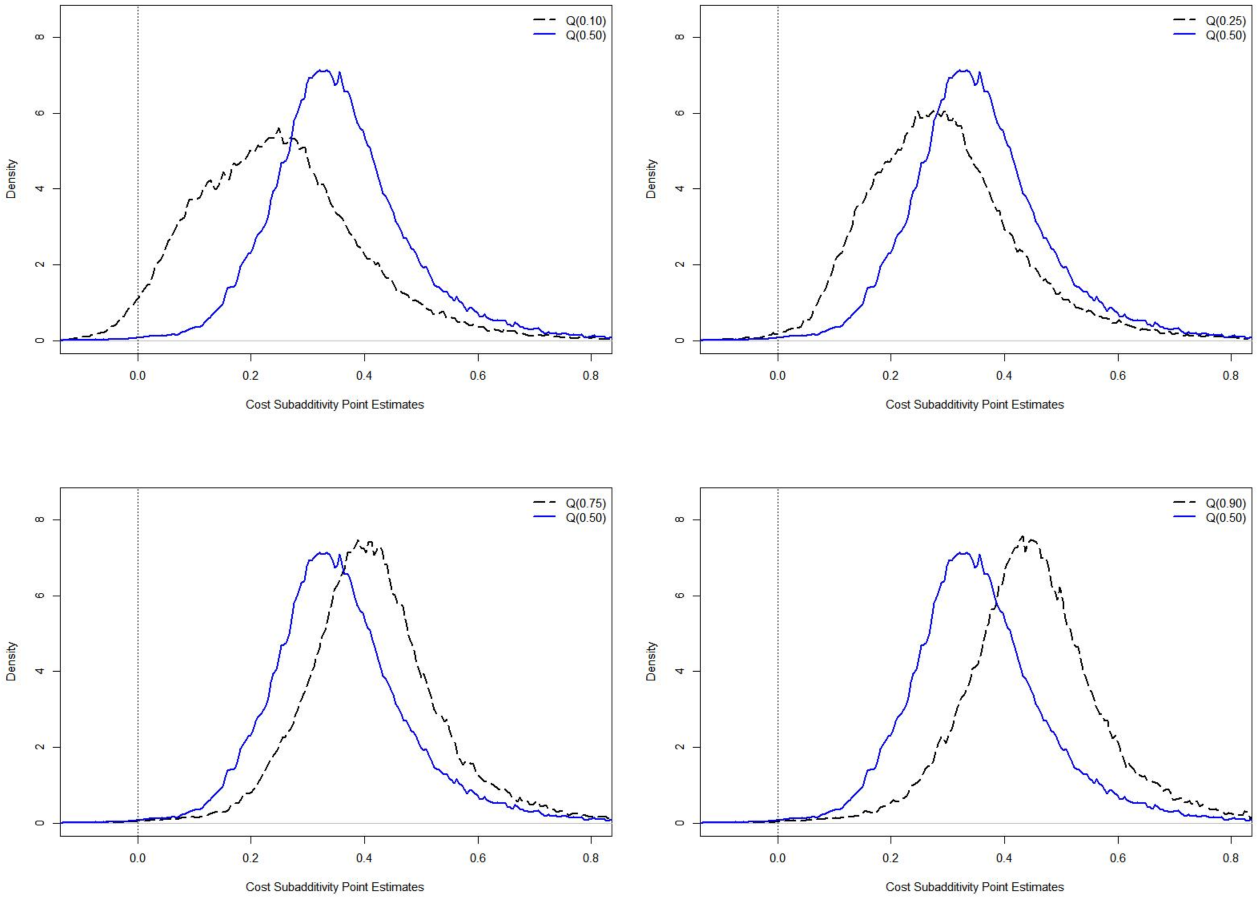

Figure 3.3. Kernel Densities of Cost Subadditivity Estimates Across Cost Quantiles

Table 3.3. p-Values for the Stochastic Dominance Tests for Scope Economies Across Cost Quantiles

	Pair-Wise				Multiple Quantiles		
	$\mathscr{Q}(0.75)$	$\mathscr{Q}(0.50)$	$\mathscr{Q}(0.25)$	$\mathscr{Q}(0.10)$	$\{\mathscr{Q}(0.75),\mathscr{Q}(0.50),\ldots\}$	$\{\mathscr{Q}(0.50),\mathscr{Q}(0.25),\ldots\}$	$\{\mathscr{Q}(0.25),\mathscr{Q}(0.10)\}$
$\mathscr{Q}(90)$	0.920	0.859	0.729	0.698	0.920	0.859	0.729
$\mathscr{Q}(75)$		0.462	0.191	0.658		0.462	0.553
$\mathscr{Q}(50)$			0.789	0.779			0.804
$\mathscr{Q}(25)$				0.975			

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 $F_\tau(\mathscr{S})$ represent the cumulative distribution functions of the $\mathscr{S}_t^*(\tau)$ 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 $\overline{\tau} \in \mathbb{T}$ with $\mathbb{T} = \{0.10, 0.25, 0.50, 0.75, 0.90\}$, we are interested in

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

We use the sub-sampling procedure suggested by Linton et al. (2005) to perform the test.[9]

The left panel of Table 3.3 reports p-values for the tests of pair-wise dominance of $\mathscr{S}_t^*(\tau)$ from the "row" quantile model over $\mathscr{S}_t^*(\tau)$ from the "column" quantile model, whereas the right panel reports p-values for the tests of dominance of $\mathscr{S}_t^*(\tau)$ from the "row" quantile over a joint multi-quantile set of $\mathscr{S}_t^*(\tau)$ 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 Figures 3.2–3.3, we can therefore conclude that banks in the higher quantiles of the cost distribution exhibit larger scope economies than do smaller banks the lower cost quantiles along the entire distribution of observable output mixes.

[9] We employ 199 equidistant sub-sample sizes $B_n = \{b_1, \ldots, b_{199}\}$, where $b_1 = [\log\log N]$, $b_{199} = [N/\log\log N]$ 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.

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

$$\mathscr{R}_t(\tau) = \left(1 - \partial \mathscr{Q}_c(\tau|\cdot)/\partial k_1\right) \Big/ \sum_m \partial \mathscr{Q}_c(\tau|\cdot)/\partial y_m, \tag{3.5.2}$$

where we replaced the usual $\log \mathscr{C}_t(\cdot)$ with the quantile function of the log-cost $\mathscr{Q}_c(\tau|\cdot)$ 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*) $\mathbb{H}_0 : \mathscr{R}_t(\tau) \leq 1$ v. $\mathbb{H}_1 : \mathscr{R}_t(\tau) > 1$ and (*ii*) $\mathbb{H}_0 : \mathscr{R}_t(\tau) = 1$ v. $\mathbb{H}_1 : \mathscr{R}_t(\tau) \neq 1$. 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 3.4 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 3.4 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 1.341 to 1.432, with banks from the higher quantiles of cost distribution exhibiting increasing returns to scale of larger magnitudes compared to those from the lower

Table 3.4. Returns to Scale Estimates

Cost		*Point Estimates*			*Inference Categories, %*			
Quantiles (τ)	Mean	1st Qu.	Median	3rd Qu.	$= \mathbf{1}$	$\neq 1$	$> \mathbf{1}$	≤ 1
$\mathscr{Q}(0.10)$	1.341	1.263	1.324	1.397	**0.88**	99.12	**99.25**	0.75
	(1.233, 1.378)	(1.228, 1.294)	(1.273, 1.357)	(1.334, 1.443)				
$\mathscr{Q}(0.25)$	1.357	1.276	1.340	1.416	**0.87**	99.13	**99.25**	0.75
	(1.305, 1.393)	(1.237, 1.306)	(1.291, 1.372)	(1.357, 1.461)				
$\mathscr{Q}(0.50)$	1.383	1.297	1.366	1.447	**0.83**	99.17	**99.28**	0.72
	(1.337, 1.425)	(1.258, 1.332)	(1.321, 1.402)	(1.392, 1.502)				
$\mathscr{Q}(0.75)$	1.411	1.32	1.394	1.481	**0.78**	99.22	**99.3**	0.7
	(1.367, 1.476)	(1.283, 1.364)	(1.351, 1.444)	(1.43, 1.548)				
$\mathscr{Q}(0.90)$	1.432	1.334	1.411	1.502	**0.74**	99.26	**99.3**	0.7
	(1.385, 1.519)	(1.293, 1.381)	(1.364, 1.466)	(1.447, 1.578)				

The left panel summarizes point estimates of $\mathscr{R}_t(\tau)$ with the corresponding two-sided 95% bias-corrected confidence intervals in parentheses. Each bank-year is classified as exhibiting IRS [$\mathscr{R}_t(\tau) > 1$] vs. non-IRS [$\mathscr{R}_t(\tau) \leq 1$] and CRS [$\mathscr{R}_t(\tau) = 1$] vs. non-CRS [$\mathscr{R}_t(\tau) \neq 1$] 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.

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-Tobòn and Kumbhakar, 2015; Malikov et al., 2015; Restrepo-Tobòn 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.

3.5.3 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

Table 3.5. Technical Change

Cost		*Point Estimates*			*Categories, %*	
Quantiles (τ)	Mean	1st Qu.	Median	3rd Qu.	$>\mathbf{0}$	≤ 0
$\mathscr{Q}(0.10)$	0.016	0.009	0.016	0.025	**73.92**	26.08
	(0.013, 0.019)	(0.006, 0.012)	(0.013, 0.021)	(0.022, 0.03)		
$\mathscr{Q}(0.25)$	0.017	0.01	0.018	0.025	**80.78**	19.22
	(0.014, 0.019)	(0.008, 0.013)	(0.013, 0.021)	(0.022, 0.028)		
$\mathscr{Q}(0.50)$	0.018	0.009	0.020	0.029	**81.56**	18.44
	(0.015, 0.02)	(0.007, 0.013)	(0.017, 0.023)	(0.025, 0.031)		
$\mathscr{Q}(0.75)$	0.019	0.007	0.023	0.033	**74.94**	25.06
	(0.017, 0.021)	(0.004, 0.01)	(0.02, 0.027)	(0.028, 0.038)		
$\mathscr{Q}(0.90)$	0.019	0.003	0.024	0.037	**71.96**	28.04
	(0.017, 0.022)	(–0.001, 0.008)	(0.02, 0.028)	(0.029, 0.042)		

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 $[TC_t(\tau) \le 0]$ using the corresponding one-sided 95% bias-corrected confidence bound. The right panel reports sample shares for each category.

measure of technological change at each cost quantile τ. Namely, from (3.3.3), we have

$$\begin{aligned} -\mathscr{T}\mathscr{C}_t(\tau) &\equiv \mathscr{Q}_c(\tau|\cdot,t) - \mathscr{Q}_c(\tau|\cdot,t-1) \\ &= \Delta L(t) + \Delta S(t) q_\tau + \\ &\quad \Big[\boldsymbol{\beta}_3 \Delta L(t) + \boldsymbol{\gamma}_3 \Delta S(t) q_\tau\Big]' \boldsymbol{v}_{it} + \frac{1}{2}\Big[\boldsymbol{\beta}_4 \Delta L(t) + \boldsymbol{\gamma}_4 \Delta S(t) q_\tau\Big]' \mathrm{vec}\big(\boldsymbol{v}_{it}\boldsymbol{v}_{it}'\big), \qquad (3.5.3) \end{aligned}$$

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

$$\begin{aligned} -TC_t(\tau) = {} & \big(\eta_\kappa + \theta_\kappa q_\tau\big) D_{\kappa,t} - \big(\eta_{\kappa-1} + \theta_{\kappa-1} q_\tau\big) D_{\kappa-1,t-1} + \\ & \Big[\big(\boldsymbol{\beta}_3 \eta_\kappa + \boldsymbol{\gamma}_3 \theta_\kappa q_\tau\big) D_{\kappa,t} - \big(\boldsymbol{\beta}_3 \eta_{\kappa-1} + \boldsymbol{\gamma}_3 \theta_{\kappa-1} q_\tau\big) D_{\kappa-1,t-1}\Big]' \boldsymbol{v}_{it} + \\ & \frac{1}{2}\Big[\big(\boldsymbol{\beta}_4 \eta_\kappa + \boldsymbol{\gamma}_4 \theta_\kappa q_\tau\big) D_{\kappa,t} - \big(\boldsymbol{\beta}_4 \eta_{\kappa-1} + \boldsymbol{\gamma}_4 \theta_{\kappa-1} q_\tau\big) D_{\kappa-1,t-1}\Big]' \mathrm{vec}\big[\boldsymbol{v}_{it}\boldsymbol{v}_{it}'\big]. \qquad (3.5.4) \end{aligned}$$

The first line in (3.5.4) 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 3.5. The right panel of the table reports results of a one-sided test of $\mathbb{H}_0 : TC_t(\tau) \le 0$ 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.

The data suggest that, in the period following the financial crisis, banks have been benefiting from the non-negligible cost-diminishing technological advances of 1.6–1.9% p.a., on average. We find that overwhelming majority of banks have experienced significant technical progress. The share of banks with statistically positive technological change estimates is at least as large as 72%, with cost diminution being slightly more prevalent among banks in the middle of the cost distribution. Overall, our results are unsurprising in light of many technological advancements that have been happening in the banking industry such as the growing networks of automated teller machines, growing credit card networks, electronic payments, internet banking, etc.; they are also consistent with earlier findings (e.g., Wheelock and Wilson, 1999; Almanidis, 2013; Malikov et al., 2015).

3.6 Conclusion

Propelled by the recent financial product innovations, banks are becoming more complex, bran-ching 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 such as investment banking, venture capital, security brokerage, insurance underwriting and asset securitization, thereby making the portfolio of products offered by pres-ent-day banks 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 comprehensively 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 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 material evidence in support of scope diseconomies.

References

Ackerberg, D. A., Benkard, C. L., Berry, S., and Pakes, A. (2007). Econometric tools for analyzing market outcomes. In Heckman, J. J. and Leamer, E. E., editors, *Handbook of Econometrics*, volume 6A. North Holland.

Ackerberg, D. A., Caves, K., and Frazer, G. (2015). Identification properties of recent production function estimators. *Econometrica*, 83(6):2411–2451.

Aitken, B., Hanson, G. H., and Harrison, A. E. (1997). Spillovers, foreign investment, and export behavior. *Journal of International Economics*, 43(1-2):103–132.

Alcácer, J. and Chung, W. (2007). Location strategies and knowledge spillovers. *Management Science*, 53(5):760–776.

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.

Almeida, P. and Kogut, B. (1999). Localization of knowledge and the mobility of engineers in regional networks. *Management Science*, 45(7):905–917.

Alvarez, R., Faruq, H., and López, R. A. (2013). Is previous export experience important for new exports? *The Journal of Development Studies*, 49(3):426–441.

Alvarez, R. and Lopez, R. A. (2005). Exporting and performance: Evidence from Chilean plants. *Canadian Journal of Economics/Revue canadienne d'économique*, 38(4):1384–1400.

Alvarez, R. and López, R. A. (2008). Is exporting a source of productivity spillovers? *Review of World Economics*, 144(4):723–749.

Amiti, M. and Konings, J. (2007). Trade liberalization, intermediate inputs, and productivity: Evidence from Indonesia. *American Economic Review*, 97(5):1611–1638.

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.

Aw, B. Y., Chung, S., and Roberts, M. J. (2000). Productivity and turnover in the export market: Micro-level evidence from the republic of Korea and Taiwan (China). *The World Bank Economic Review*, 14(1):65–90.

Aw, B.-Y. and Hwang, A. R.-m. (1995). Productivity and the export market: A firm-level analysis. *Journal of Development Economics*, 47(2):313–332.

Aw, B. Y., Roberts, M. J., and Xu, D. Y. (2011). R&D investment, exporting, and productivity dynamics. *American Economic Review*, 101(4):1312–44.

Baldwin, J. R. and Gu, W. (2003). Export-market participation and productivity performance in Canadian manufacturing. *Canadian Journal of Economics/Revue canadienne d'économique*, 36(3):634–657.

Baldwin, J. R. and Gu, W. (2004). Trade liberalization: Export-market participation, productivity growth, and innovation. *Oxford Review of Economic Policy*, 20(3):372–392.

Baltagi, B. H., Egger, P. H., and Kesina, M. (2016). Firm-level productivity spillovers in China's chemical industry: A spatial Hausman-Taylor approach. *Journal of Applied Econometrics*, 31(1):214–248.

Baltagi, B. H. and Griffin, J. M. (1988). A general index of technical change. *Journal of Political Economy*, 96(1):20–41.

Bannister, G. J. and Stolp, C. (1995). Regional concentration and efficiency in Mexican manufacturing. *European Journal of Operational Research*, 80:672–690.

Baumol, W., Panzar, J., and Willig, R. (1982). *Contestable Markets and the Theory of Industry Structure*. Harcourt, Brace & Jovanovich, San Diego.

Behrens, K., Duranton, G., and Robert-Nicoud, F. (2014). Productive cities: Sorting, seleciton, and agglomeration. *Journal of Political Economy*, 122:507–553.

Berenguer, G., Iyer, A. V., and Yadav, P. (2016). Disentangling the efficiency drivers in country-level global health programs: An empirical study. *Journal of Operations Management*, 45:30–43.

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. 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.

Bernard, A. B. and Jensen, J. B. (1999). Exceptional exporter performance: Cause, effect, or both? *Journal of International Economics*, 47(1):1–25.

Bernard, A. B. and Jensen, J. B. (2004). Why some firms export. *Review of Economics and Statistics*, 86(2):561–569.

Blalock, G. and Gertler, P. J. (2004). Learning from exporting revisited in a less developed setting. *Journal of Development Economics*, 75(2):397–416.

Blundell, R. and Bond, S. (2000). Gmm estimation with persistent panel data: An application to production functions. *Econometric Reviews*, 19(3):321–340.

Brunsdon, C. F., Fotheringham, A. S., and Charlton, M. E. (1996). Geographically weigthed regression: A method for exploring spatial nonstationarity. *Geographic Analysis*, 28:281–298.

Bustos, P. (2011). Trade liberalization, exports, and technology upgrading: Evidence on the impact of MERCOSUR on Argentinian firms. *American economic review*, 101(1):304–40.

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.

Clerides, S. K., Lach, S., and Tybout, J. R. (1998). Is learning by exporting important? micro-dynamic evidence from Colombia, Mexico, and Morocco. *The Quarterly Journal of Economics*, 113(3):903–947.

Collard-Wexler, A. and De Loecker, J. (2015). Reallocation and technology: Evidence from the US steel industry. *American Economic Review*, 105:131–171.

Combes, P.-P., Duranton, G., Gobillon, L., Puga, D., and Roux, S. (2012). The productivity advantages of large cities: Dinstinguishing agglomeration from firm selection. *Econometrica*, 80:2543–2594.

Crane, B., Albrecht, C., Duffin, K. M., and Albrecht, C. (2018). China's special economic zones: An analysis of policy to reduce regional disparities. *Regional Studies, Regional Science*, 5(1):98–107.

Dahl, M. S. and Sorenson, O. (2012). Home sweet home: Entrepreneurs' location choices and the performance of their ventures. *Management Science*, 58:1059–1071.

Das, S. R. and Nanda, A. (1999). A theory of banking structure. *Journal of Banking and Finance*, 23(6):863–895.

De Loecker, J. (2007). Do exports generate higher productivity? Evidence from Slovenia. *Journal of International Economics*, 73(1):69–98.

De Loecker, J. (2011). Product differentiation, multiproduct firms, and estimating the impact of trade liberalization on productivity. *Econometrica*, 79(5):1407–1451.

De Loecker, J. (2013). Detecting learning by exporting. *American Economic Journal: Microeconomics*, 5:1–21.

De Loecker, J., Goldberg, P. K., Khandelwal, A. K., and Pavcnik, N. (2016). Prices, markups, and trade reform. *Econometrica*, 84:445–510.

De Loecker, J. and Warzynski, F. (2012). Markups and firm-level export status. *American Economic Review*, 102(6):2437–71.

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.

Delgado, M. A., Farinas, J. C., and Ruano, S. (2002). Firm productivity and export markets: A non-parametric approach. *Journal of International Economics*, 57(2):397–422.

Demerjian, P., Lev, B., and McVay, S. (2012). Quantifying managerial ability: A new measure and validity tests. *Management Science*, 58(7):1229–1248.

DeYoung, R. and Torna, G. (2013). Nontraditional banking activities and bank failures during the financial crisis. *Journal of Financial Intermediation*, 22(3):397–421.

Doraszelski, U. and Jaumandreu, J. (2013). R&D and productivity: Estimating endogenous productivity. *Review of Economic Studies*, 80:1338–1383.

Doraszelski, U. and Jaumandreu, J. (2018). Measuring the bias of technological chance. *Journal of Political Economy*, 126:1027–1084.

Doraszelski, U. and Jaumandreu, J. (2019). Using cost minimization to estimate markups. Working Paper, University of Pennsylvania.

Duranton, G. and Puga, D. (2004). Micro-foundations of urban agglomeration economies. In Henderson, J. V. and Thisse, J. F., editors, *Handbook of Regional and Urban Economics*, volume 4. Elsevier B.V.

Efron, B. (1982). *The Jackknife, the Bootstrap, and Other Resampling Plans.* Society for Industrial and Applied Mathematics.

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association*, 82(397):171–185.

Ehrlich, I., Gallais-Hamonno, G., Liu, Z., and Lutter, R. (1994). Productivity growth and firm ownership: An analytical and empirical investigation. *Journal of Political Economy*, 102(5):1006–1038.

Elhorst, J. P. (2010). Applied spatial econometrics: Raising the bar. *Spatial Economic Analysis*, 5:9–28.

Ellison, G. and Glaeser, E. L. (1999). The geographic concentration of industry: Does natural advantage explain agglomeration? *American Economic Review*, 89(2):311–316.

Ellison, G., Glaeser, E. L., and Kerr, W. R. (2010). What causes industry agglomeration? Evidence from coagglomeration patterns. *American Economic Review*, 100(3):1195–1213.

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.

Flynn, B. B., Huo, B., and Zhao, X. (2010). The impact of supply chain integration on performance: A contingency and configuration approach. *Journal of operations management*, 28(1):58–71.

Flynn, Z., Gandhi, A., and Traina, J. (2019). Measuring markups with production data. Working Paper, University of Pennsylvania.

Fotheringham, A. S., Brunsdon, C. F., and Charlton, M. E. (2002). *Geographically Weighted Regression: The analysis of spatially varying relationships.* John Wiley & Sons, Ltd, West Sussex, England.

Gandhi, A., Navarro, S., and Rivers, D. (2020). On the identification of gross output production functions. *Journal of Political Economy*.

Gaubert, C. (2018). Firm sorting and agglomeration. *American Economic Review*, 108:3117–3153.

Glass, A. J., Kenjegaliev, A., and Kenjegalieva, K. (2020a). Spatial scale and product mix economies in U.S. banking with simultaneous spillover regimes. *European Journal of Operational Research*, 284:693–711.

Glass, A. J. and Kenjegalieva, K. (2019). A spatial productivity index in the presence of efficiency spillovers: Evidence for US banks, 1992–2015. *European Journal of Operational Research*, 273(3):1165–1179.

Glass, A. J., Kenjegalieva, K., and Douch, M. (2020b). Uncovering spatial productivity centers using asymmetric bidirectional spillovers. *European Journal of Operational Research*, 285:767–788.

Glass, A. J., Kenjegalieva, K., and Sickles, R. C. (2016a). Returns to scale and curvature in the presence of spillovers: Evidence from European countries. *Oxford Economic Papers*, 68(1):40–63.

Glass, A. J., Kenjegalieva, K., and Sickles, R. C. (2016b). A spatial autoregressive stochastic frontier model for panel data with asymmetric efficiency spillovers. *Journal of Econometrics*, 190(2):289–300.

Goldstein, S. M., Ward, P. T., Leong, G. K., and Butler, T. W. (2002). The effect of location, strategy, and operations technology on hospital performance. *Journal of Operations Mangement*, 20:63–75.

Greenaway, D. and Kneller, R. (2004). Exporting and productivity in the United Kingdom. *Oxford Review of Economic Policy*, 20(3):358–371.

Greenaway, D. and Kneller, R. (2008). Exporting, productivity and agglomeration. *European Economic Review*, 52(5):919–939.

Greenaway, D., Sousa, N., and Wakelin, K. (2004). Do domestic firms learn to export from multinationals? *European Journal of Political Economy*, 20(4):1027–1043.

Grieco, P. L. E., Li, S., and Zhang, H. (2016). Production function estimation with unobserved input price dispersion. *International Economic Review*, 57:665–689.

Grieco, P. L. E., Li, S., and Zhang, H. (2019). Input prices, productivity and trade dynamics: Longrun effects of liberalization on Chinese pain manufacturers. Working Paper, Pennsylvania State University.

Griffith, R., Redding, S., and Reenen, J. V. (2004). Mapping the two faces of R&D: Productivity growth in a panel of OECD industries. *Review of Economics and Statistics*, 86(4):883–895.

Griliches, Z. and Mairesse, J. (1998a). Production functions: The search for identification. In *In Econometrics and Economic Theory in the Twentieth Century: The Ragnar Frisch Centennial Symposium*, pages 169–203. Cambridge University Press.

Griliches, Z. and Mairesse, J. (1998b). Production functions: The search for identification. In *Econometrics and Economic Theory in the Twentieth Century: The Ragnar Frisch Centennial Symposium*, pages 169–203. Cambridge University Press.

Grossman, G. M. and Helpman, E. (2015). Globalization and growth. *American Economic Review*, 105(5):100–104.

Hall, P., Li, Q., and Racine, J. S. (2007). Nonparametric estimation of regression functions in the presence of irrelevant regressors. *Review of Economics and Statistics*, 89(4):784–789.

Hou, Z., Jin, M., and Kumbhakar, S. C. (2020). Productivity spillovers and human capital: A semiparametric varying coefficient approach. *European Journal of Operational Research*, 287:317–330.

Hsieh, C.-T. and Song, Z. M. (2015). Grasp the large, let go of the small: The transformation of the state sector in China. *Brookings Papers on Economic Activity*, (1):3.

Hu, Y., Huang, G., and Sasaki, Y. (2020). Estimating production functions with robustness against errors in the proxy variables. *Journal of Econometrics*, 215(2):375–398.

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 and 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.

Jin, M., Zhao, S., and Kumbhakar, S. C. (2019). Financial constraints and firm productivity: Evidence from Chinese manufacturing. *European Journal of Operational Research*, 275(3):1139–1156.

Jola-Sanchez, A. F., Pedraza-Martinez, A. J., Bretthauer, K. M., and Britto, R. A. (2016). Effect of armed conflicts on humanitarian operations: Total factor productivity and efficiency of rural hospitals. *Journal of Operations Management*, 45:73–85.

Kalnins, A. and Chung, W. (2004). Resource-seeking agglomeration: A study if market entry in the lodging industry. *Strategic Management Journal*, 25:689–699.

Kalnins, A. and Chung, W. (2006). Social capital, geograhy, and survival: Gujarati immigrant entrepreneurs in the U.S. lodging industry. *Management Science*, 52:233–247.

Kasahara, H. and Lapham, B. (2013). Productivity and the decision to import and export: Theory and evidence. *Journal of International Economics*, 89(2):297–316.

Keller, W. (2004). International technology diffusion. *Journal of Economic Literature*, 42(3):752–782.

Ketokivi, M., Turkulainen, V., Seppälä, T., Rouvinen, P., and Ali-Yrkkö, J. (2017). Why locate manufacturing in a high-cost country? A case study of 35 production location decisions. *Journal of Operations Management*, 49:20–30.

Koenig, P. (2009). Agglomeration and the export decisions of French firms. *Journal of Urban Economics*, 66(3):186–195.

Koenig, P., Mayneris, F., and Poncet, S. (2010). Local export spillovers in France. *European Economic Review*, 54(4):622–641.

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.

Konings, J. and Vanormelingen, S. (2015). The impact of training on productivity and wages: Firm-level evidence. *Review of Economics and Statistics*, 97:485–497.

Kulchina, E. (2016). Personal preferences, entrepreneur's location choices, and firm performance. *Management Science*, 62:1814–1829.

Kunst, R. M. and Marin, D. (1989). On exports and productivity: A causal analysis. *The Review of Economics and Statistics*, pages 699–703.

Kutlu, L., Tran, K. C., and Tsionas, M. G. (2020). A spatial stochastic frontier model with endogenous frontier and environmental variables. *European Journal of Operational Research*, 286:389–399.

Laeven, L. and Levine, R. (2007). Is there a diversification discount in financial conglomerates? *Journal of Financial Economics*, 85(2):331–367.

Lam, H. K., Yeung, A. C., and Cheng, T. E. (2016). The impact of firms' social media initiatives on operational efficiency and innovativeness. *Journal of Operations Management*, 47:28–43.

Lamarche, C. (2010). Robust penalised quantiel regression estimation for panel data. *Journal of Econometrics*, 157:396–408.

LeSage, J. and Pace, R. K. (2009). *Introduction to Spatial Econometrics.* Chapman & Hall/CRC, Boca Raton.

Levinsohn, J. and Petrin, A. (2003). Estimating production functions using inputs to control for unobservables. *The Review of Economic Studies*, 70(2):317–341.

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 & 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.

Malikov, E., Zhang, J., Zhao, S., and Kumbhakar, S. (2021). Accounting for cross-location technological heterogeneity in the measurement of operations efficiency and productivity. Working Paper, University of Nevada, Las Vegas.

Malikov, E. and Zhao, S. (2019). Cross-firm productivity spillovers in the presence of foreign investments. Working Paper, University of Nevada, Las Vegas.

Malikov, E. and Zhao, S. (2021). On the estimation of cross-firm productivity spillovers with an application to fdi. *Review of Economics and Statistics.* forthcoming.

Malikov, E., Zhao, S., and Kumbhakar, S. C. (2020). Estimation of firm-level productivity in the presence of exports: Evidence from China's manufacturing. *Journal of Applied Econometrics*, 35(4):457–480.

Mammen, E. (1993). Bootstrap and wild bootstrap for high dimensional linear models. *Annals of Statistics*, 21:255–285.

Manjón, M., Máñez, J. A., Rochina-Barrachina, M. E., and Sanchis-Llopis, J. A. (2013). Reconsidering learning by exporting. *Review of World Economics*, 149:5–22.

Markides, C. and Williamson, P. J. (1994). Related diversification, core competences and corporate performance. *Strategic Management Journal*, 15(S2):149–165.

Marshall, A. (1920). *Principle of Economics.* London: MacMillan.

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.

Moretti, E. (2004). Workers' education, spillovers, and productivity: Evidence from plant-level production functions. *American Economic Review*, 94(3):656–690.

Olley, G. S. and Pakes, A. (1996). The dynamics of productivity in the telecommunications equipment industry. *Econometrica*, 64:1263–1297.

Ouyang, D., Li, D., and Li, Q. (2006). Cross-validation and non-parametric k nearest-neighbour estimation. *Econometrics Journal*, 9:448–471.

Panzar, J. C. and Willig, R. D. (1981). Economies of scope. *American Economic Review*, 71(2):268–272.

Park, A., Yang, D., Shi, X., and Jiang, Y. (2010). Exporting and firm performance: Chinese exporters and the Asian financial crisis. *The Review of Economics and Statistics*, 92(4):822–842.

Pavcnik, N. (2002). Trade liberalization, exit, and productivity improvements: Evidence from Chilean plants. *The Review of Economic Studies*, 69(1):245–276.

Poncet, S. and Mayneris, F. (2013). French firms penetrating Asian markets: Role of export spillovers. *Journal of Economic Integration*, pages 354–374.

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-Tobòn, 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-Tobòn, 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.

Ross, A. D. and Droge, C. (2004). An analysis of operations efficiency in large-scale distribution systems. *Journal of Operations Management*, 21(6):673–688.

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.

Rothenberg, T. J. (1971). Identification in parametric models. *Econometrica*, 39:577–591.

Rumelt, R. P. (1982). Diversification strategy and profitability. *Strategic Management Journal*, 3(4):359–369.

Sala, D. and Yalcin, E. (2015). Export experience of managers and the internationalisation of firms. *The World Economy*, 38(7):1064–1089.

Salomon, R. and Jin, B. (2008). Does knowledge spill to leaders or laggards? Exploring industry heterogeneity in learning by exporting. *Journal of International Business Studies*, 39:132–150.

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.

Serpa, J. C. and Krishnan, H. (2018). The impact of supply chains on firm-level productivity. *Management Science*, 64:511–532.

Shao, J. and Tu, D. (1995). *The jackknife and bootstrap*. Springer–Verlag New York.

Singh, J. and Marx, M. (2013). Geographic constraints on knowledge spillovers: Political borders vs. spatial proximity. *Management Science*, 59(9):2056–2078.

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.

Smith, T. M. and Reece, J. S. (1999). The relationship of strategy, fit, productivity, and business performance in a services setting. *Journal of Operations Management*, 17(2):145–161.

Solow, R. M. (1957). Technical change and the aggregate production function. *Review of Economics and Statistics*, 39:312–320.

Song, Z., Storesletten, K., and Zilibotti, F. (2011). Growing like China. *American Economic Review*, 101(1):196–233.

Stiroh, K. J. (2004). Diversification in banking: Is noninterest income the answer? *Journal of Money, Credit and Banking*, pages 853–882.

Syverson, C. (2011). What determines productivity? *Journal of Economic Literature*, 49:326–365.

Tsionas, M. G. and Mallick, S. K. (2019). A bayesian semiparametric approach to stochastic frontiers and productivity. *European Journal of Operational Research*, 274(1):391–402.

Ullah, A. (1985). Specification analysis of econometric models. *Journal of Quantitative Economics*, 1:187–209.

Van Biesebroeck, J. (2005). Exporting raises productivity in sub-Saharan African manufacturing firms. *Journal of International Economics*, 67(2):373–391.

Verhoogen, E. A. (2008). Trade, quality upgrading, and wage inequality in the Mexican manufacturing sector. *The Quarterly Journal of Economics*, 123(2):489–530.

Vidoli, F. and Canello, J. (2016). Controlling for spatial heterogeneity in nonparametric efficiency models: An empirical proposal. *European Journal of Operational Research*, 249(2):771–783.

Villalonga, B. (2004). Diversification discount or premium? New evidence from the Business Information Tracking Series. *Journal of Finance*, 59(2):479–506.

Wagner, J. (2007). Exports and productivity: A survey of the evidence from firm-level data. *World*

Economy, 30(1):60–82.

Wei, Y. and Liu, X. (2006). Productivity spillovers from R&D, exports and FDI in China's manufacturing sector. *Journal of International Business Studies*, 37(4):544–557.

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.

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*.

Zhang, J. and Malikov, E. (2019). Learning by exporting and from exporters in Chilean manufacturing. Working Paper, Auburn University.

Zhang, J. and Malikov, E. (2021). Off-balance-sheet activities and scope economies in U.S. banking. Working Paper, Auburn University.

Zhao, S., Qian, B., and Kumbhakar, S. C. (2020). Estimation of productivity and markups with price dispersion: Evidence from Chinese manufacturing during economic transition. *Southern Economic Journal*.

Appendices

A Appendix to Chapter 2

A.1 The SAR Production-Function Models

A popular approach to incorporating locational/spatial effects in production models relies on spatial econometric techniques, whereby spatially-weighted averages of other firms' outputs (and sometimes inputs too) are included as additional regressors in the SAR production-function models. Not only does such a SAR specification of the production relationship continues to implausibly assume the common technology for all locations, it also becomes problematic in practice, when it comes to its estimation: (*i*) the SAR production functions imply additional, highly nonlinear parameter restrictions necessary to ensure that the conventional axioms of the production theory are not violated, albeit these are usually ignored in applied work, and (*ii*) the identification of SAR production-function models from data is hardly guaranteed due to the inapplicability of available proxy-variable estimators and the general lack of valid external instruments. In what follows, we discuss each of these considerations in detail.

Axiomatic Considerations.—For expositional simplicity, for now let us assume a deterministic log-linear (Cobb-Douglas) input-output relationship with a single input and suppress the time index. The SAR production function a là Glass et al. (2016b) in logs is given by

$$y_i = \rho \sum_{j(\neq i)=1}^{n} d_{ij} y_j + \beta_K k_i + \omega_i, \tag{A.1}$$

where $\{d_{ij} \geq 0\}$ are the non-negative spatial weights that describe the architecture of inter-firm spatial dependence. Following the convention, $d_{ii} = 0\ \forall i$ and $d_{ij} = d_{ji}\ \forall i, j$. Putting these weights for all firms together gives a symmetric $n \times n$ non-stochastic spatial weighting matrix $\mathbf{D}$ which, following the popular practice, is row-standardized.[1] Assume that $\mathbf{D}$ is known. To ensure spatial stationarity, the spatial lag parameter is $\rho \in (-1, 1)$.

The above formulation implies that the ith firm's output y_i is a function of not only its own input k_i but of all its neighbors' inputs $\{k_j\}$, and the elasticity of own capital is no longer equal to β_K. The latter are not without implications for the theoretical regularity conditions routinely assumed—explicitly or implicitly—about the production frontier.

To make matters more concrete, letting $\boldsymbol{y} = (y_1, \ldots, y_n)'$, $\boldsymbol{k} = (k_1, \ldots, k_n)'$ and $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_n)'$, which all are the $n \times 1$ vectors, we have a vector of reduced-form production functions for all firms:

$$\boldsymbol{y} = \overbrace{\left[\mathbf{I}_n - \rho\mathbf{D}\right]^{-1}}^{\mathbf{Q}} \left(\beta_K \boldsymbol{k} + \boldsymbol{\omega}\right), \tag{A.2}$$

from where we have that, for each firm i, the production function is

$$y_i = \beta_K \sum_{j=1}^{n} Q_{ij} k_j + \sum_{j=1}^{n} Q_{ij} \omega_j, \tag{A.3}$$

with Q_{ij} being the (i, j)th element of the $n \times n$ “spatial multiplier matrix” $\mathbf{Q}$. Exponentiating (A.3), we arrive at the firm i's production function in levels:

$$Y_i = K_i^{\beta_K Q_{ii}} \left[\prod_{j(\neq i)=1}^{n} K_j^{Q_{ij}} \right]^{\beta_K} \prod_{j=1}^{n} \exp\{Q_{ij}\omega_j\}. \tag{A.4}$$

Conventional production axioms imply, among other theoretical regularity conditions, the following about the shape of production function: monotonicity and concavity in inputs. Now, because the theory of production does not normally consider the possibility of inter-firm spillovers, technically it does not differentiate between the production unit's *own* and its *neighbors*' inputs. By implication however, the standard regularity conditions describe the within-unit technological relationships between the production unit's *own* inputs

[1] That is, $\sum_{j=1}^{n} d_{ij} = 1$ for all rows $i = 1, \ldots, n$, although there are alternative normalizations available.

and outputs.

It is straightforward to see that (A.4) remains monotone in K_i so long as the capital elasticity parameter is positive ($\beta_K > 0$):

$$\frac{\partial Y_i}{\partial K_i} = \beta_K Q_{ii} K_i^{\beta_K Q_{ii}-1} \left[\prod_{j(\neq i)=1}^{n} K_j^{Q_{ij}} \right]^{\beta_K} \prod_{j=1}^{n} \exp\{Q_{ij}\omega_j\} > 0 \tag{A.5}$$

because $K_i \in \Re_+$ and $Q_{ii} \in \Re_+$ for all i.

Things are not as trivial when it comes to curvature. Consider the own second partial derivative w.r.t. capital:

$$\frac{\partial^2 Y_i}{\partial K_i^2} = \beta_K Q_{ii} \left(\beta_K Q_{ii} - 1 \right) K_i^{\beta_K Q_{ii}-2} \left[\prod_{j(\neq i)=1}^{n} K_j^{Q_{ij}} \right]^{\beta_K} \prod_{j=1}^{n} \exp\{Q_{ij}\omega_j\}. \tag{A.6}$$

For (A.6) to remain non-positive in order to guarantee concavity of the production frontier, it needs be that the diagonal elements of the spatial multiplier matrix $\mathbf{Q} = \left[\mathbf{I}_n - \rho\mathbf{D}\right]^{-1}$ are such that $Q_{ii} \leq \beta_K^{-1}$ (so long as $\beta_K > 0$ just like earlier). Since Q_{ii} is a function of ρ and $\mathbf{D}$, the latter places restrictions on both the strength of spatial dependence—the magnitude of ρ—and the architecture of spatial relationships, i.e., the specification/design of $\mathbf{D}$.

This is an important implication because, depending on the value of capital elasticity β_K, these restrictions can be quite strict if the acceptable range of Q_{ii} is especially narrow. In fact, the permissible range is not $Q_{ii} \leq \beta_K^{-1}$ but

$$1 \leq Q_{ii} \leq \beta_K^{-1} \quad \forall\ i, \tag{A.7}$$

since the diagonal elements of $\mathbf{Q}$ are all no smaller than 1 by construction (see Elhorst, 2010). This already rules out the possibility of constant (internal) returns to scale commonly used in the applied productivity research, particularly at the aggregate level. Namely, if $\beta_K = 1$, it must be that $Q_{ii} = 1$ for all i which is possible only if $\rho = 0$. Consequently, cross-firm spatial spillovers in (A.1) are consistent only with the decreasing returns to scale, i.e., when $\beta_K < 1$.

The practical implementation can get complicated even when one is willing to assume the decreasing returns to scale *a priori*. Lately, when estimating production technologies from the data, it has increasingly become a norm to impose theoretical regularity conditions onto the estimand to ensure that the estimates are structurally meaningful. In the case of a SAR production function, the restriction in (A.7) would also need to be imposed. While global in nature, this constraint depends on data, is highly nonlinear and involves the inversion of an $n \times n$ matrix which can be quite computationally demanding even for moderate sample sizes (see LeSage and Pace, 2009):

$$\mathbf{i}_n \leq \operatorname{diag}\left\{\left[\mathbf{I}_n - \rho \mathbf{D}\right]^{-1}\right\} \leq \frac{1}{\beta_K}\mathbf{i}_n. \tag{A.8}$$

Identification Considerations.—A more fundamental issue with the SAR production-function models concerns their (un)identifiability from the firm-level data due to the endogeneity of the firm's allocations of own variable inputs. The latter can be equivalently characterized as the omitted variable problem, whereby variable inputs are correlated with firm unobservables which we capture using the persistent productivity term ω_{it}. Many studies considering SAR production-function models (or their Durbin extensions) leave this well-known problem unaddressed, focusing only on the endogeneity of the SAR lag term while implausibly assuming that all input regressors are exogenous or that unobservables—firm productivity, or efficiency—are purely random (see Glass et al., 2016b; Glass and Kenjegalieva, 2019; Glass et al., 2020b). Others tackle this problem under the assumption that the endogeneity-inducing correlated unobservables are time-invariant and can be controlled for via firm fixed effects (e.g., Glass et al., 2016a, 2020a).

While clearly an improvement over the random treatment of firm unobservables, the fixed-effect approach however tends to be quite unsatisfactory in practice because differencing or within-transforming of the data necessary to purge fixed effects oftentimes leaves little usable identifying variation which yields unrealistically small and statistically insignificant estimates of the capital elasticity (see Griliches and Mairesse, 1998b; Ackerberg et al., 2007; Gandhi et al., 2020). Consequently, practitioners have favored tackling endogeneity in the production-function estimation via identification schemes involving proxy variables or external instruments.

The instrument-based identification of SAR production functions has been considered by Kutlu et al. (2020), although they provide no guidance about the candidates for valid external instruments. In the stochastic-frontier productivity literature, the instrumentation is most times done using firm-level variation in prices.[2] However, the validity and practicality of using lagged firm-level prices for identification is not universal. Not only are the price data often unavailable or prone to measurement errors (Levinsohn and Petrin, 2003), but the use of prices may also be problematic on theoretical grounds (see Griliches and Mairesse, 1998b; Ackerberg et al., 2007, 2015; Flynn et al., 2019). Specifically, the validity of prices as exogenous instruments is normally justified by invoking the assumption of perfectly competitive markets. However, if firms were indeed price-takers, in theory, one should not observe the firm-level variation in prices and, without such a variation, prices cannot be used as operational instruments. Even with the aggregate prices varying exogenously across space, such a variation may be insufficient for identification as shown by Gandhi et al. (2020). If a researcher does observe the variation in prices across all individual firms, the latter variation may be reflecting differences in firms' market power and/or the quality of inputs/outputs. For instance, if the firm-level variation in input prices reflects differential quality in inputs, then random updates in prices that render prices valid instruments are likely related to productivity innovations because a more productive firm is to use more productive, higher-quality inputs (Flynn et al., 2019). Thus, be it due to the market power or quality differentials, the variation in prices will then be endogenous to firms' decisions and hence cannot help the identification (also see Gandhi et al., 2020). Furthermore, *lagging* the instruments does not help either. Putting the issue of exogeneity aside, Flynn et al. (2019) raise concerns about the strong conditions on the evolution processes that must be satisfied for the lagged prices to have any strength as instruments.[3]

The above underscores the practical appeal of proxy-variable identification strategies that do not require external instruments to identify production technologies. Despite some recent attempts at the proxy-variable identification of SAR production functions (see Hou

[2] The less common instruments include demand shifters or the external "determinants" of firm productivity/efficiency.

[3] Similar arguments can be made about the demand and productivity shifters.

et al., 2020), below we show that such proxy-variable methodologies originally designed for the estimation of non-spatial production functions generally *cannot* be extended to accommodate their SAR specifications.

For concreteness, we augment the logged SAR production function in (A.1) to include more than one input and a random shock and where we also resume time-indexing the variables:

$$y_{it} = \rho \sum_{j(\neq i)=1}^{n} d_{ij} y_{jt} + \beta_K k_{it} + \beta_M m_{it} + \omega_{it} + \eta_{it}. \tag{A.9}$$

For simplicity, here we assume that spatial relationships captured by the weights $\mathbf{D} = \{d_{ij}\}$ are time-invariant. Akin to (A.3), the corresponding reduced form of the ith firm production function at time period t is

$$y_{it} = \beta_K \sum_{j=1}^{n} Q_{ij} k_{jt} + \beta_M \sum_{j=1}^{n} Q_{ij} m_{jt} + \sum_{j=1}^{n} Q_{ij} \left[\omega_{jt} + \eta_{jt}\right], \tag{A.10}$$

where $\mathbf{Q} = \{Q_{ij}\}$ is an $n \times n$ time-invariant diagonal block of the $nT \times nT$ spatial multiplier matrix

$$\left[\mathbf{I}_{nT} - \rho \mathbf{I}_T \otimes \mathbf{D}\right]^{-1} = \mathbf{I}_T \otimes \overbrace{\left[\mathbf{I}_n - \rho \mathbf{D}\right]^{-1}}^{\mathbf{Q}}.$$

In line with the standard structural assumptions in the proxy-variable productivity literature (and to echo those we make in Section 2.2), we assume that (*i*) K_{it} is dynamically optimized with a delay and subject to the adjustment costs, whereas M_{it} is freely varying and chosen statically, (*ii*) persistent firm productivity follows a controlled (location-homogeneous) first-order Markov process with transition probability $\mathcal{P}^{\omega}(\omega_{it}|\omega_{it-1}, G_{it-1})$, and the random shock η_{it} is i.i.d., (*iii*) firms are risk-neutral and seek to maximize a discounted stream of expected life-time profits in perfectly competitive output and factor markets. Note that, in the case of a SAR formulation of spatial effects, the extent through which firm location plays a role in the production is via the SAR term $\sum_{j(\neq i)=1}^{n} d_{ij} y_{jt}$ that shifts the frontier; all production-function parameters as well as the productivity evolution process are location-invariant.

As noted earlier, in order to identify the SAR production function in (A.9), one needs to tackle the endogeneity of not only the spatial lag $\sum_{j(\neq i)=1}^{n} d_{ij} y_{jt}$ (due to the simultaneous "reflection") but also both firm inputs that are correlated with *unobservable* firm productivity ω_{it}. While the endogenous $\sum_{j(\neq i)=1}^{n} d_{ij} y_{jt}$ can be handled fairly easily using internal instruments such as the first- and higher-order spatial lags of neighbors' inputs (per the reduced form in (A.10)) as typically done in spatial models, the endogeneity of inputs requires far more finesse. Owing to the already-discussed general lack of external instruments, a popular approach to tackling this omitted variable problem is structural and relies on proxying for the "omitted" ω_{it} using the inverted material demand function. Namely, analogous to the steps we take in Section 2.3.1, making use of the Markovian nature of firm productivity, we can rewrite the SAR production function (A.9) as

$$y_{it} = \rho \sum_{j(\neq i)=1}^{n} d_{ij} y_{jt} + \beta_K k_{it} + \beta_M m_{it} + h(\omega_{it-1}, G_{it-1}) + \zeta_{it} + \eta_{it}, \tag{A.11}$$

where $h(\cdot)$ is the conditional mean of ω_{it} which, if desired, can be assumed to be linear, and ζ_{it} is a productivity innovation.

Provided that we can construct an (observable) proxy for ω_{it-1}, k_{it} and G_{it-1} are predetermined and weakly exogenous w.r.t. $\zeta_{it} + \eta_{it}$ but the freely varying m_{it} is not because it is a function of ζ_{it} (through ω_{it} based on which the firm statically chooses materials in period t). Thus, both $\sum_{j(\neq i)=1}^{n} d_{ij} y_{jt}$ and m_{it} are endogenous. Since the spatial lag is easily instrumentable, let us focus on handling the endogeneity of m_{it}.

As we explain in Section (2.3.1), in order to identify the production function in flexible inputs such as m_{it}, a solution is to exploit a structural link between the production function and the firm's (static) first-order condition for m_{it}, which is also what we do in the first step of our methodology. Thus, the firm's restricted expected profit-maximization problem w.r.t. the flexible material input subject to the already optimized dynamic input K_{it}, productivity ω_{it} and prices $(P_t^Y, P_t^M)'$ is

$$\max_{M_{it}} P_t^Y \exp\left\{\rho \sum_{j(\neq i)=1}^{n} d_{ij} \mathbb{E}[y_{jt} | \mathscr{I}_{it}]\right\} K_{it}^{\beta_K} M_{it}^{\beta_M} \exp\{\omega_{it}\} \theta - P_t^M M_{it}. \tag{A.12}$$

The optimization problem now also includes the *expected* average neighbor log-output

$\sum_{j(\neq i)=1}^{n} d_{ij}\mathbb{E}[y_{jt}|\mathscr{I}_{it}]$ net of random *ex post* shocks $\{\eta_{jt}\}$ unobservable to firms at the time of making decisions: $y_{jt} = \mathbb{E}[y_{jt}|\mathscr{I}_{it}] + \eta_{jt}\ \forall i, j$. Precisely because of the latter and so long as there are spatial spillovers across firms in that $\rho \neq 0$, the firm's first-order condition now accounts for "feedback effects" whereby the change in M_{it} affects not only Y_{it} but also—through spillovers—neighbors' outputs $\{Y_{jt}\}$ which, in turn, affect firm i's output Y_{it} again. Therefore, the corresponding first-order condition is

$$\left(\beta_M + \rho \sum_{j(\neq i)=1}^{n} d_{ij} \frac{\partial \mathbb{E}[y_{jt}|\mathscr{I}_{it}]}{\partial m_{it}}\right) P_t^Y \exp\left\{\rho \sum_{j(\neq i)=1}^{n} d_{ij}\mathbb{E}[y_{jt}|\mathscr{I}_{it}]\right\} K_{it}^{\beta_K} M_{it}^{\beta_M - 1} \exp\{\omega_{it}\}\theta = P_t^M. \tag{A.13}$$

To arrive at the material share equation, we take the log of (A.13) and subtract the production function in (A.9):

$$\begin{aligned} v_{it} &= \ln\left[\left(\beta_M + \rho \sum_{j(\neq i)=1}^{n} d_{ij} \frac{\partial \mathbb{E}[y_{jt}|\mathscr{I}_{it}]}{\partial m_{it}}\right)\theta\right] - \rho \sum_{j(\neq i)=1}^{n} d_{ij}\eta_{jt} - \eta_{it} \\ &= \ln\left[\beta_M\left(1 + \rho \sum_{j(\neq i)=1}^{n} d_{ij}Q_{ji}\right)\theta\right] \underbrace{-\rho \sum_{j(\neq i)=1}^{n} d_{ij}\eta_{jt} - \eta_{it}}_{\epsilon_{it}}, \end{aligned} \tag{A.14}$$

where we have made a substitution in the second line using the partial $\frac{\partial \mathbb{E}[y_{jt}|\mathscr{I}_{it}]}{\partial m_{it}} = Q_{ji}\beta_M$ obtained from the reduced form in (A.10) and, just like before, $v_{it} = \ln\left(P_t^M M_{it}\right) - \ln\left(P_t^Y Y_{it}\right)$ is the (observable) log nominal share of material costs in total revenue. Note that, unlike in a model without spatial lag, the composite error term $\epsilon_{it} \equiv -\left(\rho \sum_{j(\neq i)=1}^{n} d_{ij}\eta_{jt} + \eta_{it}\right)$ in (A.14) follows a spatial moving average process. Although ϵ_{it} is spatially correlated, this has no impact on identification because, by assumption, shocks $\{\eta_{it}\}$ are all i.i.d. Further note that each Q_{ji} element in (A.14) is a function of spatial weights and the still-unknown ρ. To make this more explicit, we write (A.14) as

$$v_{it} = \ln\left[\beta_M\left(1 + \rho\mathbf{D}_{(i)}\left(\left[\mathbf{I}_n - \rho\mathbf{D}\right]^{-1}\right)_{(j)}\right)\theta\right] + \epsilon_{it}, \tag{A.15}$$

where $\mathbf{A}_{(i)}$ and $\mathbf{A}_{(j)}$ respectively denote the ith row and jth column of some matrix $\mathbf{A}$ (and recall that $d_{ij} = 0\ \forall i = j$).

The material share equation (A.14) is a nonlinear regression containing *no* endogenous covariates. It therefore might appear at first that it can be seamlessly estimated via nonlinear least squares. However, the parameters in this regression $\boldsymbol{\alpha} \equiv (\beta_M, \rho, \theta)'$ are *not* identified. To make this unidentification more apparent, we recast the nonlinear least-squares estimator of (A.15) in a GMM framework.

Namely, we write the nonlinear equation (A.15) as $v_{it} = h_{it}(\mathbf{D}, \boldsymbol{\alpha}) + \epsilon_{it}$, where $h_{it}(\cdot)$ is the regression function. Consider now the identification of $\boldsymbol{\alpha}$ in the following just-identified nonlinear GMM problem that is equivalent to the nonlinear least-squares estimation:

$$\boldsymbol{\alpha}_0 = \arg\min_{\boldsymbol{\alpha}} \mathbb{E}\left[\frac{\partial h_{it}(\mathbf{D}, \boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}}\Big(v_{it} - h_{it}(\mathbf{D}, \boldsymbol{\alpha})\Big)\right]' \mathbf{W} \mathbb{E}\left[\frac{\partial h_{it}(\mathbf{D}, \boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}}\Big(v_{it} - h_{it}(\mathbf{D}, \boldsymbol{\alpha})\Big)\right], \qquad \text{(A.16)}$$

where

$$\frac{\partial h_{it}(\mathbf{D}, \boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} = \begin{bmatrix} {\beta_M}^{-1} \\ \frac{1}{1+\rho \mathbf{D}_{(i)}\left(\left[\mathbf{I}_n - \rho \mathbf{D}\right]^{-1}\right)_{(j)}} \left(\mathbf{D}_{(i)}\left(\left[\mathbf{I}_n - \rho \mathbf{D}\right]^{-1}\right)_{(j)} + \rho \mathbf{D}_{(i)}\left(\left[\mathbf{I}_n - \rho \mathbf{D}\right]^{-1} \mathbf{D}\left[\mathbf{I}_n - \rho \mathbf{D}\right]^{-1}\right)_{(j)}\right) \\ \theta^{-1} \end{bmatrix}, \qquad \text{(A.17)}$$

and $\mathbf{W}$ is a symmetric positive-definite moment-weighting matrix.

To see that (A.16)–(A.17) do *not* identify all of the $\boldsymbol{\alpha}$ parameters, letting the element in the second row of $\frac{\partial h_{it}(\mathbf{D}, \boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}}$ be pictorially denoted by "□," consider the corresponding information matrix:

$$\Psi(\boldsymbol{\alpha}) = \mathbb{E}\left[\frac{\partial h_{it}(\mathbf{D}, \boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} \frac{\partial h_{it}(\mathbf{D}, \boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}'}\right] = \mathbb{E}\begin{bmatrix} \frac{1}{\beta_M^2} & \frac{1}{\beta_M}\square & \frac{1}{\beta_M \theta} \\ \frac{1}{\beta_M}\square & \square^2 & \frac{1}{\theta}\square \\ \frac{1}{\beta_M \theta} & \frac{1}{\theta}\square & \frac{1}{\theta^2} \end{bmatrix}. \qquad \text{(A.18)}$$

The above 3 × 3 matrix $\Psi(\boldsymbol{\alpha})$ has rank of 1. Thus, the information matrix for the GMM problem in (A.16) when evaluated at the true parameter values $\Psi(\boldsymbol{\alpha}_0)$ will be rank-deficient, and the parameters in $\boldsymbol{\alpha}$ will therefore be unidentified (see Rothenberg, 1971).

It is also important to note that augmenting the nonlinear least-squares moment restrictions with the unconditional moment corresponding to θ given in (2.3.8) analogously to what we do in the first step of our identification methodology will *not* remedy the unidentification problem. More concretely, recalling that $\theta \equiv \mathbb{E}\left[\exp\left\{\eta_{it}\right\}\right]$ and inverting the composite error ϵ_{it} appearing in (A.15) from the spatial moving average process that it follows, we have the additional GMM moment restriction:

$$\begin{aligned} 0 = \mathbb{E}\left[g_{it}(\mathbf{D},\boldsymbol{\alpha})\right] &\equiv \mathbb{E}\left[\exp\left\{\eta_{it}\right\} - \theta\right] \\ &= \mathbb{E}\left[\exp\left\{-\left(\left[\mathbf{I}_n + \rho\mathbf{D}\right]^{-1}\right)_{(i)} \boldsymbol{\epsilon}_t\right\} - \theta\right], \end{aligned} \tag{A.19}$$

where $\boldsymbol{\epsilon}_t = (\epsilon_{1t},\ldots,\epsilon_{nt})'$ and each $\epsilon_{it} = v_{it} - h_{it}(\mathbf{D},\boldsymbol{\alpha}) = v_{it} - \ln\left[\beta_M\left(1 + \rho\mathbf{D}_{(i)}\left(\left[\mathbf{I}_n - \rho\mathbf{D}\right]^{-1}\right)_{(j)}\right)\theta\right]$. With this, let us consider the information matrix $I(\boldsymbol{\alpha}) = [\Psi(\boldsymbol{\alpha}),\Phi(\boldsymbol{\alpha})]$ for an augmented GMM problem, where

$$\Phi(\boldsymbol{\alpha}) = \mathbb{E}\left[\frac{\partial g_{it}(\mathbf{D},\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}}\right] = \mathbb{E}\begin{bmatrix} \beta_M^{-1}\exp\left\{-\left(\left[\mathbf{I}_n + \rho\mathbf{D}\right]^{-1}\right)_{(i)}\boldsymbol{\epsilon}_t\right\}\left(\left[\mathbf{I}_n + \rho\mathbf{D}\right]^{-1}\right)_{(i)}\mathbf{i}_n \\ \frac{\partial g_{it}(\mathbf{D},\boldsymbol{\alpha})}{\partial \rho} \\ \theta^{-1}\exp\left\{-\left(\left[\mathbf{I}_n + \rho\mathbf{D}\right]^{-1}\right)_{(i)}\boldsymbol{\epsilon}_t\right\}\left(\left[\mathbf{I}_n + \rho\mathbf{D}\right]^{-1}\right)_{(i)}\mathbf{i}_n - 1 \end{bmatrix}. \tag{A.20}$$

It is apparent that the information matrix $I(\boldsymbol{\alpha})$ even when including the additional moment is still not full-rank, and the material share equation remains unidentified.

We have thus shown that, notwithstanding the Gandhi et al. (2020) result for non-spatial proxy-variable estimators, the SAR production function cannot be identified in flexible inputs by exploiting a structural link between the production function and the firm's (static) optimality conditions. Consequently, the whole model is unidentified. This is in stark contrast with the main result of our paper. Our ability to capture locational effects in production and achieve proxy-variable identification of the production technology and firm productivity in the presence of technology spillovers and agglomeration economies stems from our fundamentally different conceptualization of cross-firm spatial interdependence. Our methodology incorporates firm location through local smoothing, which models the production technology for each location as the geographically weighted average of the input-output *relationships* for firms in the nearby locations, whereas the SAR production-function models formulate locational aspects using the spatially weighed averages of the output/inputs *quantities* while keeping the production technology location-invariant.

A.2 Translog Technology

Our methodology can adopt more flexible specifications of the firm's production technology. The log-quadratic translog specification provides a natural extension of the log-linear Cobb-Douglas form. See De Loecker and Warzynski (2012) and De Loecker et al. (2016) for recent applications of the translog production functions in the structural proxy estimation. Just like we have done in (2.3.1), we generalize the standard fixed-parameter translog specification to accommodate potential locational heterogeneity in production by letting its coefficients vary with the firm's location in a nonparametric way, i.e.,

$$\begin{aligned}\ln F_{|S_i}(\cdot) = {} & \beta_K(S_i)k_{it} + \tfrac{1}{2}\beta_{KK}(S_i)k_{it}^2 + \beta_L(S_i)l_{it} + \tfrac{1}{2}\beta_{LL}(S_i)l_{it}^2 + \beta_M(S_i)m_{it} + \tfrac{1}{2}\beta_{MM}(S_i)m_{it}^2 + \\ & \beta_{KL}(S_i)k_{it}l_{it} + \beta_{KM}(S_i)k_{it}m_{it} + \beta_{LM}(S_i)l_{it}m_{it}. \qquad \text{(A.21)}\end{aligned}$$

Our methodology can then be modified as follows.

First step.—The firm's first-order condition for the static optimization in (2.2.4) with respect to M_{it} is now takes the following form:

$$\ln P_t^Y + \ln F_{|S_i} + \ln\left[\beta_M(S_i) + \beta_{MM}(S_i)m_{it} + \beta_{KM}(S_i)k_{it} + \beta_{LM}(S_i)l_{it}\right] - m_{it} + \omega_{it} + \ln\theta = \ln P_t^M, \qquad \text{(A.22)}$$

where $\ln F_{|S_i}$ equals the semiparametric translog technology in (A.21). The location-specific material share equation corresponding to this optimality condition is now given by

$$v_{it} = \ln\left(\left[\beta_M(S_i) + \beta_{MM}(S_i)m_{it} + \beta_{KM}(S_i)k_{it} + \beta_{LM}(S_i)l_{it}\right]\theta\right) - \eta_{it}, \qquad \text{(A.23)}$$

where $\beta_M(S_i) + \beta_{MM}(S_i)m_{it} + \beta_{KM}(S_i)k_{it} + \beta_{LM}(S_i)l_{it}$ is the material elasticity function. Analogous to the discussion in Section 2.3.1, the above log material share equation identifies the locationally varying material-related production-function parameters ($\beta_M(S_i)$, $\beta_{MM}(S_i)$,

$\beta_{KM}(S_i), \beta_{LM}(S_i))'$ as well as the mean of exponentiated shocks θ based on the mean-orthogonality condition $\mathbb{E}[\eta_{it}|\mathscr{I}_{it}] = \mathbb{E}[\eta_{it}|k_{it}, l_{it}, m_{it}, S_i] = \mathbb{E}[\eta_{it}] = 0$.

Second step.—Having identified the production technology in the dimension of its endogenous freely varying input M_{it}, we can focus on the rest of production function. With the already identified $y^*_{it} \equiv y_{it} - \beta_M(S_i)m_{it} - \frac{1}{2}\beta_{MM}(S_i)m^2_{it} - \beta_{KM}(S_i)k_{it}m_{it} - \beta_{LM}(S_i)l_{it}m_{it}$ and using the inverted conditional material demand derived from (A.22) to substitute for ω_{it-1}, we now have the analogue of (2.3.11):

$$\begin{aligned} y^*_{it} = {} & \beta_K(S_i)k_{it} + \tfrac{1}{2}\beta_{KK}(S_i)k^2_{it} + \beta_L(S_i)l_{it} + \tfrac{1}{2}\beta_{LL}(S_i)l^2_{it} + \beta_{KL}(S_i)k_{it}l_{it} + \rho_0(S_i) + \\ & \rho_1(S_i)\Big[v^*_{it-1} - \beta_K(S_i)k_{i,t-1} - \tfrac{1}{2}\beta_{KK}(S_i)k^2_{i,t-1} - \beta_L(S_i)l_{i,t-1} - \tfrac{1}{2}\beta_{LL}(S_i)l^2_{i,t-1} - \\ & \beta_{KL}(S_i)k_{i,t-1}l_{i,t-1}\Big] + \rho_2(S_i)G_{it-1} + \zeta_{it} + \eta_{it}, \end{aligned} \tag{A.24}$$

where

$$\begin{aligned} v^*_{it-1} = {} & \ln[P^M_{t-1}/P^Y_{t-1}] - \ln([\beta_M(S_i) + \beta_{MM}(S_i)m_{it-1} + \beta_{KM}(S_i)k_{it-1} + \beta_{LM}(S_i)l_{it-1}]\theta) + \\ & [1 - \beta_M(S_i)]m_{it-1} - \tfrac{1}{2}\beta_{MM}(S_i)m^2_{it-1} - \beta_{KM}(S_i)k_{it-1}m_{it-1} - \beta_{LM}(S_i)l_{it-1}m_{it-1} \end{aligned}$$

is already identified/observable and predetermined with respect to $\zeta_{it} + \eta_{it}$. All covariates in (A.24) are predetermined and can self-instrument thereby identifying the translog model.

Estimation.—The estimation methodology here mirrors that used for the Cobb-Douglas model except that the local-constant least-squares estimator in the first step is now also nonlinear.

Assuming that all unknown locationally varying coefficient functions are smooth and twice continuously differentiable in the neighborhood of $S_i = s$, the location-specific material share equation in (A.23) can be locally approximated around s via constants as

$$\begin{aligned} v_{it} & \approx \ln\Big(\Big[\beta_M(s) + \beta_{MM}(s)m_{it} + \beta_{KM}(s)k_{it} + \beta_{LM}(s)l_{it}\Big]\theta\Big) - \eta_{it} \\ & \approx \ln\Big([\beta_M(s)\theta] + [\beta_{MM}(s)\theta]m_{it} + [\beta_{KM}(s)\theta]k_{it} + [\beta_{LM}(s)\theta]l_{it}\Big) - \eta_{it}, \end{aligned} \tag{A.25}$$

with the corresponding kernel estimator of $\Theta_1(s) = [\beta_M(s)\theta, \beta_{MM}(s)\theta, \beta_{KM}(s)\theta, \beta_{LM}(s)\theta]'$ being

$$\widehat{\Theta}_1(s) = \underset{\Theta_1(s)}{\arg\min} \sum_i \sum_t \mathscr{K}_{h_1}(S_i, s)\Big(v_{it} - \ln\Big([\beta_M(s)\theta] + [\beta_{MM}(s)\theta]m_{it} + [\beta_{KM}(s)\theta]k_{it} + [\beta_{LM}(s)\theta]l_{it}\Big)\Big)^2. \tag{A.26}$$

To estimate the material elasticity parameter functions $[\beta_M(s), \beta_{MM}(s), \beta_{KM}(s), \beta_{LM}(s)]'$ around $S_i = s$ net of constant θ, we first recover $\widehat{\theta}$ as

$$\widehat{\theta} = \sum_i \sum_t \exp\Big\{\ln\Big([\widehat{\beta_M(s)\theta}] + [\widehat{\beta_{MM}(s)\theta}]m_{it} + [\widehat{\beta_{KM}(s)\theta}]k_{it} + [\widehat{\beta_{LM}(s)\theta}]l_{it}\Big) - v_{it}\Big\} \tag{A.27}$$

and then use it to scale $\widehat{\Theta}_1(s)$ which yields our first-step estimator:

$$\big[\widehat{\beta}_M(s), \widehat{\beta}_{MM}(s), \widehat{\beta}_{KM}(s), \widehat{\beta}_{LM}(s)\big]' = nT\widehat{\Theta}_1(s)\Big/\widehat{\theta}. \tag{A.28}$$

Using these first-step local estimates, we then construct $\widehat{y}^*_{it} = y_{it} - \widehat{\beta}_M(S_i)m_{it} - \frac{1}{2}\widehat{\beta}_{MM}(S_i)m^2_{it} - \widehat{\beta}_{KM}(S_i)k_{it}m_{it} - \widehat{\beta}_{LM}(S_i)l_{it}m_{it}$ and $\widehat{v}^*_{it-1} = \ln[P^M_{t-1}/P^Y_{t-1}] - \ln([\widehat{\beta}_M(S_i) + \beta_{MM}(S_i)m_{it-1} + \widehat{\beta}_{KM}(S_i)k_{it-1} + \widehat{\beta}_{LM}(S_i)l_{it-1}]\theta) + [1 - \widehat{\beta}_M(S_i)]m_{it-1} - \frac{1}{2}\widehat{\beta}_{MM}(S_i)m^2_{it-1} - \widehat{\beta}_{KM}(S_i)k_{it-1}m_{it-1} - \widehat{\beta}_{LM}(S_i)l_{it-1}m_{it-1}$.

Analogous to the first-step estimation, we then locally approximate each unknown parameter function in (A.24) via local-constant approach. Collectively denoting all unknown parameters in the equation as $\Theta_2(S_i) = [\beta_K(S_i), \beta_{KK}(S_i), \beta_L(S_i), \beta_{LL}(S_i), \beta_{KL}(S_i), \rho_0(S_i), \rho_1(S_i), \rho_2(S_i)]'$, the second-step local-constant nonlinear least-squares estimator in the neighborhood of $S_i = s$ is then given by

$$\begin{aligned}\widehat{\Theta}_2(s) = \underset{\Theta_2(s)}{\arg\min} \sum_i \sum_t \mathscr{K}_{h_2}(S_i, s)\Big(&y^*_{it} - \beta_K(S_i)k_{it} - \tfrac{1}{2}\beta_{KK}(S_i)k^2_{it} - \beta_L(S_i)l_{it} - \tfrac{1}{2}\beta_{LL}(S_i)l^2_{it} - \\ &\beta_{KL}(S_i)k_{it}l_{it} - \rho_1(S_i)\Big[v^*_{it-1} - \beta_K(S_i)k_{i,t-1} - \tfrac{1}{2}\beta_{KK}(S_i)k^2_{i,t-1} - \beta_L(S_i)l_{i,t-1} - \\ &\tfrac{1}{2}\beta_{LL}(S_i)l^2_{i,t-1} - \beta_{KL}(S_i)k_{i,t-1}l_{i,t-1}\Big] - \rho_0(S_i) - \rho_2(S_i)G_{it-1}\Big)^2.\end{aligned} \tag{A.29}$$

A.3 Finite-Sample Performance of the Estimator

We investigate the finite-sample performance of our proposed estimation procedure in a set of Monte Carlo experiments. Our data generating process (DGP) builds on the setup in Grieco et al. (2016) and Malikov et al. (2020) which we modify to allow for locational heterogeneity.

Without loss of generality, we dispense with labor and consider the production process with two inputs only: a quasi-fixed capital and freely varying materials. Let the true technology take a semiparametric locationally-varying Cobb-Douglas form:

$$Y_{it} = K_{it}^{\beta_K(S_i)} M_{it}^{\beta_M(S_i)} \exp\{\omega_{it}\} \exp\{\eta_{it}\}. \tag{A.30}$$

To simplify matters, we assume that all firms are located on a straight line, with the univariate location variable S_i indexing their relative location. We assume a discrete uniform spatial distribution of firms, with $S_i \in \mathbb{S} = \{0.50, 0.51, \ldots, 0.98, 0.99\}$ and $D = 50$ locations. The random disturbance is $\eta_{it} \sim$ i.i.d. $\mathbb{N}(0, 0.07^2)$.

We assume the decreasing returns to scale across all locations but let the scale elasticity differ across locations. Concretely, we have that the returns to scale increase as one moves rightwards in space $\mathbb{S}$ by having the two input elasticity functions smoothly vary across locations as follows:

$$\beta_K(S_i) = 0.2 + 0.1 S_i \tag{A.31}$$

$$\beta_M(S_i) = 0.4 + 0.1 \exp(S_i^2). \tag{A.32}$$

The productivity components are generated as follows. We model the persistent productivity as a location-specific exogenous AR(1) process:

$$\omega_{it} = \rho_0(S_i) + \rho_1(S_i)\omega_{it-1} + \zeta_{it}, \tag{A.33}$$

where we set $\rho_0(S_i) = 0.5 + S_i - S_i^2$ and $\rho_1(S_i) = 0.7 \ \forall\ S_i$. In this, we assume that the mean firm

Table A.1. Second-Step Estimates of Locationally-Varying Parameters

	Panel A: Kernel-Smoothing			Panel B: Sample-Splitting		
	$\beta_K(\cdot)$	$\rho_1(\cdot)$	$\rho_2(\cdot)$	$\beta_K(\cdot)$	$\rho_1(\cdot)$	$\rho_2(\cdot)$
			$n = 100$			
Mean Bias	−0.0113	0.0065	0.0046	−0.0059	−0.0231	−0.0297
RMSE	0.0569	0.0731	0.0329	0.1725	0.2552	0.1152
MAE	0.0449	0.0593	0.0267	0.1446	0.1991	0.0885
			$n = 200$			
Mean Bias	−0.0069	0.0037	0.0036	−0.0117	0.0016	−0.0085
RMSE	0.0388	0.0508	0.0238	0.1196	0.1577	0.0693
MAE	0.0311	0.0413	0.0193	0.0937	0.1239	0.0546
			$n = 400$			
Mean Bias	−0.0070	0.0030	0.0053	−0.0073	0.0034	−0.0028
RMSE	0.0278	0.0356	0.0178	0.0797	0.1027	0.0456
MAE	0.0223	0.0290	0.0144	0.0608	0.0808	0.0362

Ours is a kernel-smoothing estimator which uses information from *all* locations, albeit weighting it based on the proximity to a location of interest. The sample-splitting estimator is essentially a "frequency estimator" which splits the data sample by location to estimates location-specific parameters using information from that location only. $T = 10$ throughout.

productivity is the highest in the middle of $\mathbb{S}$ and symmetrically diminishes in both direction therefrom. The firm's initial level of productivity ω_{i1} is set to $\rho_0(S_i)$ and therefore is determined purely by the firm's location. The productivity innovation is $\zeta_{it} \sim$ i.i.d. $\mathbb{N}(0, 0.04^2)$.

The firm's capital is set to evolve according to $K_{it} = I_{it-1} + (1-\delta_i)K_{it-1}$, with the firm-specific depreciation rates δ_i uniformly drawn from $\{0.05, 0.075, 0.10, 0.125, 0.15\}$. The initial levels of capital K_{i0} is drawn from $\mathbb{U}(10, 200)$ identically and independently distributed over i. The investment function takes the following form: $I_{it-1} = K_{it-1}^{\alpha_1} \exp\{\alpha_2 \omega_{it-1}\}$, where $\alpha_1 = 0.8$ and $\alpha_2 = 0.1$.

The materials M_{it} series is generated solving the firm's restricted expected profit maximization problem along the lines of (2.2.4). The conditional demand for M_{it} is given by

$$M_{it} = \arg\max_{\mathscr{M}_{it}} \left\{P_t^Y K_{it}^{\beta_K(S_i)} \mathscr{M}_{it}^{\beta_M(S_i)} \exp\{\omega_{it}\}\theta - P_t^M \mathscr{M}_{it}\right\} = \left[\beta_M(S_i) K_{it}^{\beta_K(S_i)} \exp\{\omega_{it}\}\right]^{1/(1-\beta_M(S_i))}, \tag{A.34}$$

where, in the second equality, we have normalized $P_t^M = \theta \;\forall\; t$ and have assumed no temporal variation in output prices: $P_t^Y = 1$ for all t.

We estimate the model via the two-step kernel-smoothing estimation algorithm outlined in Section 2.3.2. Although we cross-validate the optimal number of nearest neighbors (h) in the empirical application, to conserve computational time, in our simulations we rely on the result that the optimal h when cross-validating is $h \propto n^{4/(4+\dim(S_i))}$ (see Ouyang et al., 2006) and set $h = 0.3(nT)^{4/5}$. We consider a balanced panel of $n = \{100, 200, 400\}$ firms operating during $T = 10$ periods. Each panel is simulated $Q = 500$ times. For each simulation repetition, we compute the mean bias, the root mean squared error (RMSE) and the mean absolute error (MAE) over all firms. Panel A of Table A.1 reports these metrics averaged across Q simulations for the capital elasticity $\beta_K(S_i)$ and the productivity parameters $\rho_0(S_i)$ and $\rho_1(S_i)$.[4]

The simulation results for our estimator are encouraging and show that our methodology recovers the true parameters remarkably well, thereby lending strong support to the validity of our identification strategy. As expected of a consistent estimator, the estimation becomes more stable as n grows. Furthermore, our estimator significantly outperforms a crude—albeit computationally simpler—alternative estimator which splits the data sample by location to (parametrically) estimates location-specific parameters using the information from that location only. The results for this sample-splitting estimator are summarized in Panel B of Table A.1. Such an alternative estimation procedure is also less practical because its feasibility is dependent on having enough data from each unique location. Our estimation procedure is immune to this problem because it uses information from *all* locations but with varying degree of relative importance as determined by their proximity to a location of interest.

A.4 Bootstrap Inference

Due to a multi-step nature of our estimator as well as the presence of nonparametric components, computation of the asymptotic variance of the estimators is not simple. For statistical inference, we therefore use bootstrap. We approximate sampling distributions of the

[4] We omit the results corresponding to the material elasticity $\beta_M(S_i)$ estimated in the first step because the estimator yields very precise estimates even for small sample sizes.

estimators via wild residual block bootstrap that takes into account a panel structure of the data, with all the steps bootstrapped jointly owing to a sequential nature of our estimation procedure. Concretely, the bootstrap algorithm is as follows.

1. Compute the two steps of our estimation procedure using the original data. Denote the obtained estimates as $\left(\widehat{\beta_M}(S_i), \widehat{\theta}, \widehat{\Theta}(S_i)'\right)'$ for all $i = 1,\ldots,n$. Let the (negative of) first-step residuals be $\{\widehat{\eta}_{it}\}$ and the second-step residuals be $\{\widehat{\zeta_{it}+\eta_{it}}\}$. Recenter these.

2. Generate bootstrap weights ξ_i^b for all cross-sectional units $i = 1,\ldots,n$ from the Mammen (1993) two-point mass distribution:

$$\xi_i^b = \begin{cases} \frac{1+\sqrt{5}}{2} & \text{with prob.} \quad \frac{\sqrt{5}-1}{2\sqrt{5}} \\ \frac{1-\sqrt{5}}{2} & \text{with prob.} \quad \frac{\sqrt{5}+1}{2\sqrt{5}}. \end{cases} \tag{A.35}$$

   Next, for each observation (i,t) with $i = 1,\ldots,n$ and $t = 1,\ldots,T$, jointly generate a new bootstrap first-step disturbance $\eta_{it}^b = \xi_i^b \times \widehat{\eta}_{it}$ and a new bootstrap second-step disturbance $(\zeta_{it}+\eta_{it})^b = \xi_i^b \times (\widehat{\zeta_{it}+\eta_{it}})$.

3. Generate a new bootstrap first-step outcome variable via $v_{it}^b = \ln\left[\widehat{\beta}_M(S_i)\widehat{\theta}\right] - \eta_{it}^b$ for all $i = 1,\ldots,n$ and $t = 1,\ldots,T$.

4. Generate a new bootstrap second-step outcome variable using $y_{it}^{*b} = \widehat{\beta}_K(S_i)k_{it} + \widehat{\beta}_L(S_i)l_{it} + \widehat{\rho}_0(S_i) + \widehat{\rho}_1(S_i)\left[\widehat{\nu}_{it-1}^* - \widehat{\beta}_K(S_i)k_{it-1} - \widehat{\beta}_L(S_i)l_{it-1}\right] + \widehat{\rho}_2(S_i)G_{it-1} + (\zeta_{it}+\eta_{it})^b$ for all $i = 1,\ldots,n$ and $t = 1,\ldots,T$, where $\widehat{\nu}_{it-1}^* = \ln[P_{t-1}^M/P_{t-1}^Y] - \ln[\widehat{\beta}_M(S_i)\widehat{\theta}] + [1-\widehat{\beta}_M(S_i)]m_{it-1}$ is constructed using the original parameter estimates.

5. Recompute the first step using $\{v_{it}^b\}$ in place of $\{v_{it}\}$ and, for all i, denote the obtained coefficient estimates as $\left(\widehat{\beta}_M^b(S_i), \widehat{\theta}^b\right)'$. Use these bootstrap estimates to construct $\widehat{\nu}_{it-1}^{*b} = \ln[P_{t-1}^M/P_{t-1}^Y] - \ln[\widehat{\beta}_M^b(S_i)\widehat{\theta}^b] + [1-\widehat{\beta}_M^b(S_i)]m_{it-1}$ for all $i = 1,\ldots,n$ and $t = 1,\ldots,T$.

6. Recompute the second step using $\{y_{it}^{*b}\}$ in place of $\{\widehat{y}_{it}^*\}$. When recomputing the models, also use $\{\widehat{\nu}_{it-1}^{*b}\}$ in place of $\{\widehat{\nu}_{it-1}^*\}$. For all $i = 1,\ldots,n$, denote the obtained coefficient estimates as $\widehat{\Theta}^b(S_i) = \left(\widehat{\beta}_K^b(S_i), \widehat{\beta}_L^b(S_i), \widehat{\rho}_0^{\omega b}(S_i), \widehat{\rho}_1^{\omega b}(S_i), \widehat{\rho}_2^{\omega b}(S_i)\right)'$.

Table A.2. Coverage Probability of the Two-Sided 95% Bootstrap Confidence Interval

	$\overline{\beta}_K$	$\beta_K(0.65)$	$\beta_K(0.75)$	$\beta_K(0.85)$
Panel A: True Parameters are Location-Specific				
$n = 200$	0.873	0.923	0.901	0.850
$n = 400$	0.843	0.907	0.933	0.883
$n = 800$	0.950	0.933	0.943	0.953
Panel B: True Parameters are Location-Invariant				
$n = 200$	0.937	0.943	0.943	0.950
$n = 400$	0.963	0.957	0.970	0.960
$n = 800$	0.963	0.967	0.963	0.977

Panel A (B) corresponds to the DGP with locationally-varying (location-invariant) coefficients. In both cases, the estimation was performed using our kernel-weighting estimator under the presumption that coefficients are location-specific. $T = 10$ throughout.

7. Repeat steps 2 through 6 of the algorithm B times.

Inference is performed using the bootstrap percentile confidence intervals. Let the observation-specific estimand of interest be denoted by $\mathscr{E}$, e.g., the firm i's labor elasticity coefficient $\beta_L(S_i)$ or the returns to scale defined as the sums of $\beta_K(S_i)$, $\beta_L(S_i)$ and $\beta_M(S_i)$. To test two-tailed hypotheses, we can use the empirical distribution of $\{\widehat{\mathscr{E}}^1,\ldots,\widehat{\mathscr{E}}^B\}$ to estimate the *two*-sided $(1-\alpha)\times 100\%$ confidence bounds for $\mathscr{E}$ as an interval between the $[\alpha/2\times 100]$th and $[(1-\alpha/2)\times 100]$th percentiles of the bootstrap distribution. Naturally, for one-tailed hypotheses, to estimate the *one*-sided lower or upper $(1-\alpha)\times 100\%$ confidence bound, we can use the $[\alpha\times 100]$th or $[(1-\alpha)\times 100]$th bootstrap percentiles, respectively.

Finite-Sample Performance.—Using the DGP described in Appendix A.3, we investigate a finite-sample performance of the above bootstrap procedure in a simulation. This is of interest because of the complexity of our multi-step estimation procedure, which makes establishing the validity of bootstrap nontrivial. We focus on the two-sided 95% confidence intervals ($\alpha = 0.05$) for (i) the *average* elasticity of capital across all locations $\overline{\beta}_K = \frac{1}{D}\sum_{S_i\in\mathbb{S}}\beta_K(S_i)$ and (ii) capital elasticity at select locations $\beta_K(S_0)$. We choose $S_0 = \{0.65, 0.75, 0.85\}$ which roughly correspond to quartiles of $\mathbb{S}$. To conserve computational time, the number of simulations is $Q = 300$ with $B = 200$ bootstrap replications per each simulation.

Panel A in Table A.2 reports coverage probabilities of the bootstrap confidence intervals for capital elasticity under our DGP for the sample size $n = \{200, 400, 800\}$. The cover-

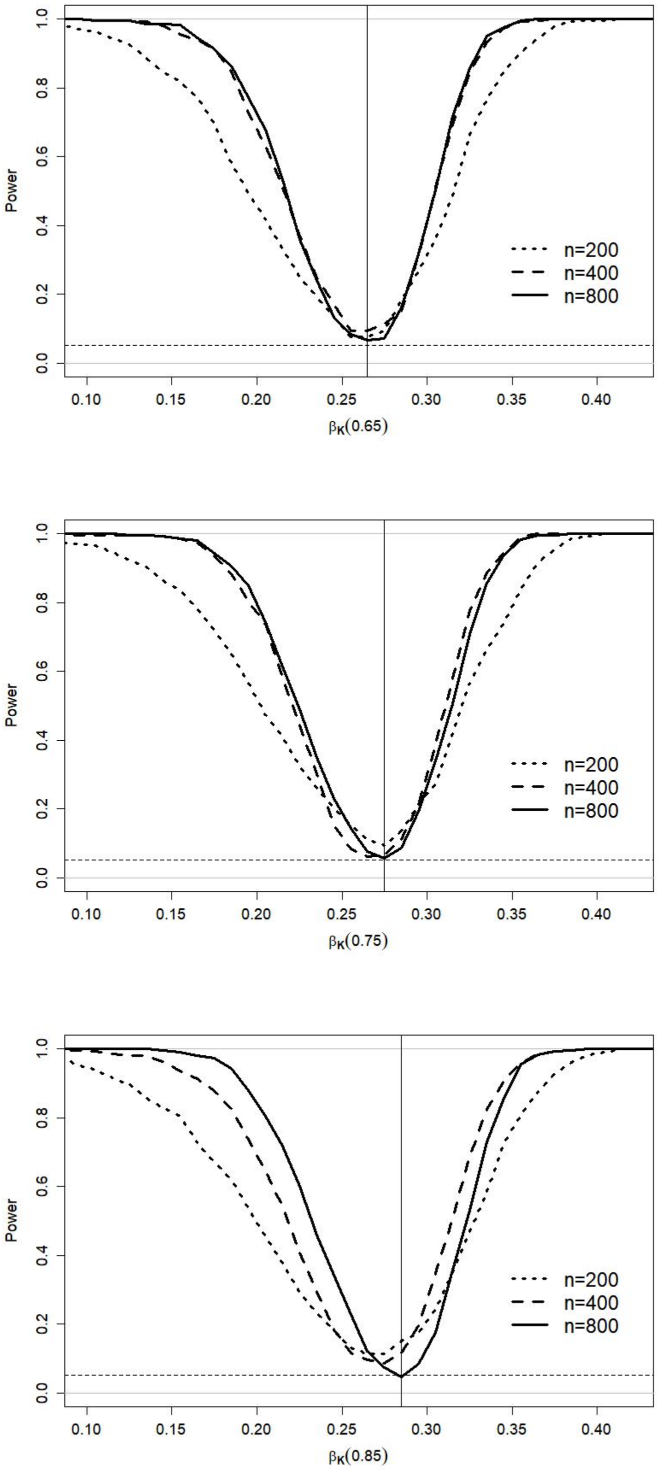

Figure A.1. Power of the Two-Sided 95% Bootstrap Confidence Interval

(Vertical line corresponds to the true $\beta_K(s)$ value) age probability is estimated as the relative frequency (over Q simulations) of the estimated 95% confidence interval containing the true elasticity value. We also plot power curves for these confidence intervals; see Figure A.1. Here, power is computed as the relative rejection frequency against different nulls on the x-axis. The simulations show a satisfactory performance of our bootstrap confidence intervals in finite samples. The results indicate that there may be size distortions for small n, which is common for nonparametric tests. However, for a sample size modestly large enough, the estimated coverage is close to the correct coverage. The intervals exhibit good power, which improves as n grows as anticipated of a consistent test.

When the DGP is such that the true parameters are actually location-invariant (i.e., fixed), the performance of bootstrap improves and the coverage probabilities match the nominal confidence level even for small n. See Panel B of Table A.2. This is expected because the nonparametric estimation improves significantly when the true process is parametric.

Bias-Corrected Inference.—Efron's (1982) bias-corrected bootstrap percentile confidence intervals provide means to robustify inference by correcting for the estimator's finite-sample bias. In this case, the bias-corrected *two*-sided $(1-\alpha)\times 100\%$ confidence bounds for $\mathscr{E}$ are estimated as an interval between the $[a_1 \times 100]$th and $[a_2 \times 100]$th percentiles of the bootstrap distribution, where $a_1 = \Phi\left(2\widehat{z}_0 + \Phi^{-1}(\alpha/2)\right)$ and $a_2 = \Phi\left(2\widehat{z}_0 + \Phi^{-1}(1-\alpha/2)\right)$ with $\Phi(\cdot)$ being the standard normal cdf along with its quantile function $\Phi^{-1}(\cdot)$. Parameter $\widehat{z}_0 = \Phi^{-1}\left(\#\left\{\widehat{\mathscr{E}}^b < \widehat{\mathscr{E}}\right\}/B\right)$ is a bias-correction factor measuring median bias, with $\#\{\mathscr{A}\}$ being a count function that returns the number of times event $\mathscr{A}$ is true. Analogously, to estimate the *one*-sided lower or upper $(1-\alpha)\times 100\%$ confidence bound with bias correction, we can respectively use the $[o_1 \times 100]$th or $[o_2 \times 100]$th percentiles of the bootstrap distribution, where $o_1 = \Phi\left(2\widehat{z}_0 + \Phi^{-1}(\alpha)\right)$ and $o_2 = \Phi\left(2\widehat{z}_0 + \Phi^{-1}(1-\alpha)\right)$. Note that the bias-corrected confidence interval need not contain the point estimate if the finite-sample bias is large.

A.5 Specification Test of Location Invariance

Given that our semiparametric locationally varying production model nests a more traditional fixed-parameter specification that implies locational invariance/homogeneity of the production function and the productivity evolution as a special case, we can formally discriminate between the two models to see if the data support our more flexible modeling approach.

More concretely, we test the null hypothesis of a production model in which the technology is *common* to firms across all locations:

$$\ln F_{|S_i}(\cdot) = \ln F(\cdot) = \beta_K k_{it} + \beta_L l_{it} + \beta_M m_{it}, \tag{A.36}$$

and firm productivity evolves according to a location-homogeneous first-order Markov process with

$$h^{\omega}_{|S_i}(\cdot) = h^{\omega}(\cdot) = \rho_0 + \rho_1 \omega_{it-1} + \rho_2 G_{it-1}. \tag{A.37}$$

The location-invariant fixed-coefficient analogue of our model under the null of locational homogeneity is therefore given by

$$y_{it} = \beta_K k_{it} + \beta_L l_{it} + \beta_M m_{it} + \rho_0 + \rho_1 \omega_{it-1} + \rho_2 G_{it-1} + \zeta_{it} + \eta_{it}, \tag{A.38}$$

which we test against our semiparametric varying-coefficient alternative in (2.3.4). This is, essentially, the test of overall relevancy of S_i.

To test this hypothesis, we use Ullah's (1985) nonparametric goodness-of-fit test based on the comparison of the restricted (under H_0) and unrestricted (under H_1) models. First, let the estimator under H_0 be denoted by "tilde" whereas the estimator under H_1 be denoted by "hat." Then, the residual-based test statistic is $T_n = (RSS_0 - RSS_1)/RSS_1$, where $RSS_0 = \sum_i \sum_t (\widetilde{\zeta_{it} + \eta_{it}})^2$ and $RSS_1 = \sum_i \sum_t (\widehat{\zeta_{it} + \eta_{it}})^2$ are respectively the second-step residual sum of squares under the (restricted parametric) null and the (unrestricted semiparametric) alternative.[5] Intuitively, the test statistic is expected to converge to zero under the

[5] We use the *second*-step residuals because they already incorporate information about the first-step estima-

null and is positive under the alternative; hence the test is one-sided. To approximate the null distribution of T_n, we use wild panel-data block-bootstrap by resampling residuals from the model under the null.

To approximate the null distribution of T_n, we use wild panel-data block-bootstrap by resampling residuals from model under the null. The algorithm builds on that described in Appendix A.4.

1. Using the original data, compute the two steps of both the locationally-invariant (under H_0) and locationally varying (under H_1) models. Denote the estimates from the restricted model as $\left[\widetilde{\beta}_M, \widetilde{\theta}, \widetilde{\Theta}'\right]'$ and the estimates from the unrestricted alternative as $\left[\widehat{\beta}_M(S_i), \widehat{\theta}, \widehat{\Theta}(S_i)'\right]'$ for all $i = 1, \ldots, n$. Let the (negative of) first-step residuals under the null be $\{\widetilde{\eta}_{it}\}$ and those under the alternative be $\{\widehat{\eta}_{it}\}$. Also, obtain the second-step residuals under the null $\{\widetilde{\zeta_{it} + \eta_{it}}\}$ and under the alternative $\{\widehat{\zeta_{it} + \eta_{it}}\}$. Use the latter to compute the test statistic T_n.
2. Generate bootstrap weights ξ_i^b for all cross-sectional units $i = 1, \ldots, n$ from the Mammen (1993) two-point mass distribution in (A.35). Next, for each observation (i, t) with $i = 1, \ldots, n$ and $t = 1, \ldots, T$, generate a new bootstrap first-step disturbance $\eta_{it}^b = \xi_i^b \times \widetilde{\eta}_{it}$ and a new bootstrap second-step disturbance $(\zeta_{it} + \eta_{it})^b = \xi_i^b \times (\widetilde{\zeta_{it} + \eta_{it}})$. When constructing these bootstrap disturbances, use re-centered residuals.
3. Generate a new bootstrap first-step outcome variable based on the specification under H_0. From the first step, we have $v_{it}^b = \ln[\widetilde{\beta}_M \widetilde{\theta}] - \eta_{it}^b$ for all $i = 1, \ldots, n$ and $t = 1, \ldots, T$.
4. Recompute the first step of both the locationally-invariant and locationally varying models using $\{v_{it}^b\}$ in place of $\{v_{it}\}$ and denote the obtained parameter estimates as $\left[\widetilde{\beta}_M^b, \widetilde{\theta}^b\right]$ under the null and as $\left[\widehat{\beta}_M^b(S_i), \widehat{\theta}^b\right]$ under the alternative.
5. Generate a new bootstrap second-step outcome variable based on the specification under H_0. From the second step, we have $y_{it}^{*b} = \widetilde{\beta}_K k_{it} + \widetilde{\beta}_L l_{it} + \widetilde{\rho}_0^{\omega} + \widetilde{\rho}_1^{\omega}\left[\widetilde{v}_{it-1}^* - \widetilde{\beta}_K k_{it-1} - \widetilde{\beta}_L l_{it-1}\right] + \widetilde{\rho}_2^{\omega} G_{it-1} + (\zeta_{it} + \eta_{it})^b$ for all $i = 1, \ldots, n$ and $t = 1, \ldots, T$, where $\widetilde{v}_{it-1}^* = \ln[P_{t-1}^M / P_{t-1}^Y] - \ln[\widetilde{\beta}_M \widetilde{\theta}] + [1 - \widetilde{\beta}_M] m_{it-1}$ is constructed using the original parameter estimates.

tion by virtue of sequential construction.

6. Recompute the second step of both models using $\{y_{it}^{*b}\}$ in place of $\{y_{it}^{*}\}$. When recomputing the models, also use $\widetilde{v}_{it-1}^{*b}$ in place of $\widetilde{v}_{it-1}^{*}$ for the restricted model and $\widehat{v}_{it-1}^{*b}$ in place of $\widehat{v}_{it-1}^{*}$ for the unrestricted model, where $\widetilde{v}_{it-1}^{*b} = \ln[P_{t-1}^{M}/P_{t-1}^{Y}] - \ln[\widetilde{\beta}_{M}^{b}\widetilde{\theta}^{b}] + [1 - \widetilde{\beta}_{M}^{b}]m_{it-1}$ and $\widehat{v}_{it-1}^{*b} = \ln[P_{t-1}^{M}/P_{t-1}^{Y}] - \ln[\widehat{\beta}_{M}^{b}(S_i)\widehat{\theta}^{b}] + [1 - \widehat{\beta}_{M}^{b}(S_i)]m_{it-1}$ are constructed using the bootstrap parameter estimates. Denote the obtained parameter estimates as $\widetilde{\Theta}^{b} = [\widetilde{\beta}_{K}^{b}, \widetilde{\beta}_{L}^{b}, \widetilde{\rho}_{0}^{\omega b}, \widetilde{\rho}_{1}^{\omega b}, \widetilde{\rho}_{2}^{\omega b}]'$ and $\widehat{\Theta}^{b}(S_i) = [\widehat{\beta}_{K}^{b}(S_i), \widehat{\beta}_{L}^{b}(S_i), \widehat{\rho}_{0}^{\omega b}(S_i), \widehat{\rho}_{1}^{\omega b}(S_i), \widehat{\rho}_{2}^{\omega b}(S_i)]'$.

7. Recompute the second-step bootstrap residuals under the null $\{\widetilde{\zeta_{it} + \eta_{it}}\}^{b} = y_{it} - \widetilde{\beta}_{K}^{b}k_{it} - \widetilde{\beta}_{L}^{b}l_{it} - \widetilde{\beta}_{M}^{b}m_{it} - \widetilde{\rho}_{0}^{\omega b} - \widetilde{\rho}_{1}^{\omega b}\left[\widehat{v}_{it-1}^{*b} - \widetilde{\beta}_{K}^{b}k_{it-1} - \widetilde{\beta}_{L}^{b}l_{it-1}\right] - \widetilde{\rho}_{2}^{\omega b}G_{it-1}$ and under the alternative $\{\widehat{\zeta_{it} + \eta_{it}}\}^{b} = y_{it} - \widehat{\beta}_{K}^{b}(S_i)k_{it} - \widehat{\beta}_{L}^{b}(S_i)l_{it} - \widehat{\beta}_{M}^{b}(S_i)m_{it} - \widehat{\rho}_{0}^{\omega b}(S_i) - \widehat{\rho}_{1}^{\omega b}(S_i)\left[\widehat{v}_{it-1}^{*b} - \widehat{\beta}_{K}^{b}(S_i)k_{it-1} - \widehat{\beta}_{L}^{b}(S_i)l_{it-1}\right] - \widehat{\rho}_{2}^{\omega b}(S_i)G_{it-1}$. Use these to compute the bootstrap test statistic T_n^b.

8. Repeat steps 2 through 7 of the algorithm B times.

Use the empirical distribution of $B+1$ bootstrap statistics $\{T_n^b\}$, where the first bootstrap replica is the test statistic T_n calculated from the original data in Step 1, to obtain the p-value as $\sum_b \mathbb{1}\{T_n^b \geq T_n\}/(B+1)$. In our empirical application, the number of bootstrap iteration is set to $B = 999$.

Estimating the Location-Invariant Model.—The location-invariant model specified in (A.36)–(A.37) is fully parametric and a special case of our locationally-varying model when $S_i = S_0$ for all i. It is therefore can be estimated following our methodology in (2.3.14)–(2.3.18) but by letting the adaptive bandwidths in both steps [$R_{h_1}(s)$ and $R_{h_2}(s)$] diverge to ∞ which would, in effect, obviate the need to locally weight the data because all kernels will be the same. We can then set $\mathscr{K}_{h_1}(S_i, s) = \mathscr{K}_{h_2}(S_i, s) = 1$ for all i.

When the production technology and the productivity process are global and do not vary across locations as described in (A.36)–(A.37), the location-invariant analogue of the first-step material share equation in (2.3.6) is given by

$$v_{it} = \ln[\beta_M \theta] - \eta_{it}, \tag{A.39}$$

and that of the second-step proxied production function in (2.3.11) is given by

$$y_{it}^{*} = \beta_K k_{it} + \beta_L l_{it} + \rho_0 + \rho_1 \left[v_{it-1}^{*} - \beta_K k_{it-1} - \beta_L l_{it-1} \right] + \rho_2 G_{it-1} + \zeta_{it} + \eta_{it}, \tag{A.40}$$

where $v_{it-1}^{*} = \ln[P_{t-1}^{M} / P_{t-1}^{Y}] - \ln[\beta_M \theta] + [1 - \beta_M] m_{it-1}$.

To estimate the material elasticity, denoting the unknown $\ln[\beta_M \theta]$ as some constant b_M, we have the following counterpart of the estimator in (2.3.14) which is just a sample mean:

$$\widehat{b}_M = \frac{1}{nT} \sum_i \sum_t v_{it}. \tag{A.41}$$

With it, we obtain the counterpart of (2.3.17) that estimates β_M:

$$\begin{aligned}
\widehat{\beta}_M &= nT \exp\{\widehat{b}_M\} \Big/ \sum_i \sum_t \exp\{\widehat{b}_M - v_{it}\} \\
&= nT \exp\left\{ \frac{1}{nT} \sum_i \sum_t v_{it} \right\} \Big/ \sum_i \sum_t \exp\left\{ \frac{1}{nT} \sum_i \sum_t v_{it} - v_{it} \right\}.
\end{aligned} \tag{A.42}$$

Using $\widehat{y}_{it}^{*} \equiv y_{it} - \widehat{\beta}_M m_{it}$ and $\widehat{v}_{it-1}^{*} = \ln[P_{t-1}^{M} / P_{t-1}^{Y}] - \ln[\widehat{\beta}_M \theta] + [1 - \widehat{\beta}_M] m_{it-1}$, we arrive at the location-invariant counterpart of the second-step estimator in (2.3.19):

$$\widehat{\Theta} = \arg\min_{\Theta} \sum_i \sum_t \left(\widehat{y}_{it}^{*} - \beta_K k_{it} - \beta_L l_{it} - \rho_0 - \rho_1 \left[\widehat{v}_{it-1}^{*} - \beta_K k_{it-1} - \beta_L l_{it-1} \right] + \rho_2 G_{it-1} \right)^2, \tag{A.43}$$

where $\Theta = [\beta_K, \beta_L, \rho_0, \rho_1, \rho_2]'$ and which can be estimated via the usual nonlinear least squares.

For inference, we follow the same bootstrap steps as those for our main model in Appendix A.4 except that the estimated location-invariant fixed coefficients are used in place of the locationally-varying coefficients.

B Appendix to Chapter 3

B.1 Cost Elasticities

Tables B.1-B.2 summarize point estimates of cost elasticities along with their corresponding two-sided bias-corrected 95% confidence intervals in parentheses.

Table B.1. Cost Elasticity Estimates

Variables	Mean	1st Qu.	Median	3rd Qu.	Mean	1st Qu.	Median	3rd Qu.
	Cost Quantile: $\mathcal{Q}(10)$				**Cost Quantile:** $\mathcal{Q}(25)$			
Y_1	0.509	0.458	0.522	0.574	0.502	0.452	0.515	0.566
	(0.493, 0.532)	(0.443, 0.486)	(0.506, 0.545)	(0.554, 0.601)	(0.488, 0.524)	(0.438, 0.479)	(0.5, 0.537)	(0.548, 0.591)
Y_2	0.091	0.074	0.091	0.11	0.09	0.073	0.09	0.109
	(0.083, 0.1)	(0.066, 0.081)	(0.082, 0.1)	(0.1, 0.118)	(0.082, 0.098)	(0.065, 0.08)	(0.081, 0.099)	(0.098, 0.117)
Y_3	0.049	0.028	0.047	0.067	0.051	0.03	0.049	0.07
	(0.041, 0.054)	(0.022, 0.032)	(0.039, 0.051)	(0.058, 0.073)	(0.044, 0.056)	(0.023, 0.034)	(0.041, 0.053)	(0.06, 0.075)
K_1	0.141	0.097	0.132	0.175	0.139	0.096	0.13	0.171
	(0.123, 0.162)	(0.076, 0.116)	(0.115, 0.153)	(0.155, 0.199)	(0.121, 0.159)	(0.076, 0.115)	(0.111, 0.15)	(0.151, 0.192)
K_2	0.012	0.005	0.013	0.02	0.013	0.006	0.015	0.022
	(0.007, 0.015)	(0.001, 0.008)	(0.009, 0.016)	(0.016, 0.023)	(0.009, 0.016)	(0.002, 0.009)	(0.011, 0.017)	(0.017, 0.024)
W_1	0.078	0.064	0.08	0.094	0.076	0.063	0.077	0.089
	(0.064, 0.089)	(0.055, 0.074)	(0.064, 0.093)	(0.073, 0.111)	(0.061, 0.086)	(0.054, 0.073)	(0.061, 0.088)	(0.07, 0.103)
W_3	0.175	0.135	0.167	0.213	0.172	0.13	0.164	0.21
	(0.162, 0.189)	(0.124, 0.146)	(0.154, 0.181)	(0.197, 0.231)	(0.161, 0.186)	(0.122, 0.141)	(0.152, 0.177)	(0.197, 0.227)
	Cost Quantile: $\mathcal{Q}(50)$				**Cost Quantile:** $\mathcal{Q}(75)$			
Y_1	0.492	0.443	0.505	0.555	0.482	0.433	0.494	0.543
	(0.475, 0.512)	(0.428, 0.465)	(0.487, 0.525)	(0.533, 0.578)	(0.462, 0.501)	(0.415, 0.452)	(0.473, 0.514)	(0.517, 0.567)
Y_2	0.087	0.071	0.088	0.106	0.085	0.069	0.086	0.104
	(0.078, 0.094)	(0.063, 0.078)	(0.078, 0.096)	(0.096, 0.114)	(0.075, 0.091)	(0.06, 0.076)	(0.076, 0.094)	(0.093, 0.112)
Y_3	0.055	0.033	0.053	0.074	0.058	0.036	0.057	0.079
	(0.049, 0.06)	(0.028, 0.037)	(0.047, 0.057)	(0.066, 0.081)	(0.053, 0.066)	(0.032, 0.041)	(0.051, 0.064)	(0.072, 0.089)
K_1	0.135	0.095	0.127	0.165	0.131	0.093	0.123	0.159
	(0.115, 0.152)	(0.075, 0.114)	(0.106, 0.144)	(0.138, 0.184)	(0.107, 0.147)	(0.074, 0.112)	(0.1, 0.139)	(0.129, 0.175)
K_2	0.015	0.007	0.017	0.024	0.017	0.009	0.019	0.027
	(0.012, 0.018)	(0.004, 0.011)	(0.014, 0.02)	(0.02, 0.027)	(0.014, 0.021)	(0.005, 0.012)	(0.016, 0.023)	(0.023, 0.032)
W_1	0.072	0.062	0.072	0.082	0.067	0.059	0.068	0.077
	(0.058, 0.081)	(0.051, 0.072)	(0.057, 0.083)	(0.065, 0.093)	(0.054, 0.077)	(0.045, 0.068)	(0.053, 0.078)	(0.062, 0.086)
W_3	0.167	0.124	0.158	0.206	0.162	0.117	0.152	0.202
	(0.157, 0.178)	(0.115, 0.134)	(0.145, 0.168)	(0.192, 0.219)	(0.148, 0.172)	(0.104, 0.126)	(0.136, 0.161)	(0.184, 0.213)

Table B.2. Cost Elasticity Estimates (cont.)

Variables	Mean	1st Qu.	Median	3rd Qu.
		Cost Quantile: $\mathscr{Q}(90)$		
Y_1	0.476	0.427	0.487	0.536
	(0.455, 0.497)	(0.408, 0.447)	(0.464, 0.509)	(0.507, 0.561)
Y_2	0.083	0.067	0.085	0.103
	(0.073, 0.09)	(0.057, 0.074)	(0.074, 0.093)	(0.092, 0.111)
Y_3	0.06	0.037	0.059	0.082
	(0.055, 0.068)	(0.033, 0.043)	(0.053, 0.068)	(0.073, 0.093)
K_1	0.128	0.092	0.121	0.156
	(0.104, 0.144)	(0.072, 0.11)	(0.098, 0.137)	(0.129, 0.171)
K_2	0.018	0.009	0.02	0.029
	(0.015, 0.022)	(0.006, 0.013)	(0.017, 0.024)	(0.024, 0.033)
W_1	0.065	0.056	0.065	0.074
	(0.051, 0.074)	(0.041, 0.066)	(0.051, 0.075)	(0.061, 0.083)
W_3	0.159	0.112	0.148	0.2
	(0.143, 0.171)	(0.099, 0.126)	(0.132, 0.16)	(0.18, 0.212)

B.2 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 $\left[\widehat{\boldsymbol{\eta}}', \widehat{\beta}_0, \widehat{\boldsymbol{\beta}_1}', \widehat{\boldsymbol{\beta}}_2', \widehat{\boldsymbol{\beta}}_1^{*\prime}, \widehat{\boldsymbol{\beta}}_2^{*\prime}\right]'$, location fixed effects $\{\widehat{\lambda}_i\}$ and residuals $\{\widehat{u}_{it}\}$.

(ii) Generate bootstrap weights ω_i^b for each cross-section/bank i from the two-point mass distribution:

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

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

(iii) Construct a new bootstrap outcome variable: $c_{it}^b = \widehat{\beta}_0 + \sum_\kappa \widehat{\eta}_\kappa D_{\kappa,t} + \left[\widehat{\boldsymbol{\beta}}_1 + \widehat{\boldsymbol{\beta}}_1^* \sum_\kappa \widehat{\eta}_\kappa D_{\kappa,t}\right]' \boldsymbol{v}_{it} + \frac{1}{2}\left[\widehat{\boldsymbol{\beta}}_2 + \widehat{\boldsymbol{\beta}}_2^* \sum_\kappa \widehat{\eta}_\kappa D_{\kappa,t}\right]' \mathrm{vec}\left(\boldsymbol{v}_{it}\boldsymbol{v}_{it}'\right) + \widehat{\lambda}_i + u_{it}^b$ for all $i = 1,\dots,n$ and $t = 1,\dots,T$.

(iv) Recompute the Step 1 estimators in (3.3.8)–(3.3.10) 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 $\widehat{u}_{it}^{b} = c_{it} - \widehat{\beta}_{0}^{b} - \sum_{\kappa} \widehat{\eta}_{\kappa}^{b} D_{\kappa,t} - \left[\widehat{\boldsymbol{\beta}}_{1}^{b} + \widehat{\boldsymbol{\beta}}_{1}^{*b} \sum_{\kappa} \widehat{\eta}_{\kappa}^{b} D_{\kappa,t}\right]' \boldsymbol{v}_{it} - \frac{1}{2}\left[\widehat{\boldsymbol{\beta}}_{2}^{b} + \widehat{\boldsymbol{\beta}}_{2}^{*b} \sum_{\kappa} \widehat{\eta}_{\kappa}^{b} D_{\kappa,t}\right]' \mathrm{vec}\left[\boldsymbol{v}_{it}\boldsymbol{v}_{it}'\right] - \widehat{\lambda}_{i}^{b}$.

(v) Reestimate the Step 2 estimator in (3.3.14) and the Step 3 estimator in (3.3.18) using $\widehat{u}_{it}^{b}$ in place of $\widehat{u}_{it}$ 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 $\mathscr{S}_{t}^{*}$) 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 $\widehat{\mathscr{E}}$. We can use the empirical distribution of $\{\widehat{\mathscr{E}}^{1},\ldots,\widehat{\mathscr{E}}^{B}\}$ to estimate the bias-corrected *two*-sided $(1-\alpha)\times 100\%$ confidence bounds for $\mathscr{E}$ as an interval between the $[a_1 \times 100]$th and $[(1-a_2)\times 100]$th percentiles of the bootstrap distribution, where $a_1 = \Phi\left(2\widehat{z}_0 + \Phi^{-1}(\alpha/2)\right)$ and $a_2 = \Phi\left(2\widehat{z}_0 + \Phi^{-1}(1-\alpha/2)\right)$ with $\Phi(\cdot)$ being the standard normal cdf along with its quantile function $\Phi^{-1}(\cdot)$. Parameter $\widehat{z}_0 = \Phi^{-1}\left(\#\{\widehat{\mathscr{E}}^{b} < \widehat{\mathscr{E}}\}/B\right)$ is a bias-correction factor measuring median bias, with $\#\{\mathscr{A}\}$ being a count function that returns the number of times event $\mathscr{A}$ 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 $o_1 = \Phi\left(2\widehat{z}_0 + \Phi^{-1}(\alpha)\right)$ and $o_2 = \Phi\left(2\widehat{z}_0 + \Phi^{-1}(1-\alpha)\right)$.