
Article:
Glass, A.J. and Kenjegalieva, K. (2023) Dynamic returns to scale and geography in U.S. banking. Papers in Regional Science, 102 (1). pp. 53-86. ISSN: 1056-8190

https://doi.org/10.1111/pirs.12713

Received: 11 June 2022 | Revised: 3 October 2022 | Accepted: 21 December 2022

DOI: 10.1111/pirs.12713

FULL ARTICLE

Dynamic returns to scale and geography in U.S. banking

Anthony J. Glass [1] | Karligash Kenjegalieva [2]

[1]Sheffield University Management School, Conduit Road, Sheffield, S10 1FL, UK

[2]School of Business and Economics and Centre for Productivity and Performance, Loughborough University, Loughborough, Leics, UK

Correspondence
Karligash Kenjegalieva, School of Business and Economics and Centre for Productivity and Performance, Loughborough University, Loughborough, Leics, UK.
Email: k.a.kenjegalieva@lboro.ac.uk

Abstract

We observe spatial cost dependence among medium-sized and large U.S. banks (1998Q1–2020Q4). We contribute to the literature by accounting for this using an accessible dynamic spatial econometric cost model. For a movement along a bank's output expansion path, we calculate the cost returns that spillover to/from the bank. The noticeable impacts of the 2020 COVID pandemic are on the spillover cost returns and not the own returns. These spillover returns suggest the pandemic led to the smallest (largest) banks becoming suboptimally smaller (bigger). A number of banks with high-ranking spillover returns have geographically concentrated branches and/or specialize in particular activities.

KEYWORDS
branch networks, overlapping geographical operations, bank cost interdependence, spill-ins and spill-outs

JEL CLASSIFICATION
C23, D24, G21, R10

1 | INTRODUCTION

There are marked differences between the domestic geography of the branch networks of a number of large U.S. banks. A notable difference is the wide variation in the number of states where large banks operate. Based on

the 215 large U.S. banks in 2020 that are part of our full sample, the average state coverage of their branch networks is 4.6 states, which is an increase from an average of 1.4 states in 1998. This coverage in 2020 ranges from banks with no interstate branching to a bank with branches in 40 states. In particular, 59 of these large banks operate in a single state, 129 operate in three or fewer states, 48 have branches in more than five states, and five have highly geographically diverse branch networks covering more than 20 states. The total number of branches these banks operate increased from 14,600 in 1998 to 51,605 in 2020, with a peak of 52,140 in 2013. Over the same period, the largest four banks accounted for 24 − 27% of the total number of branches and the average state coverage of their branch networks ranged from 7.3 states in 1998 to 32.5 in 2020. Such differences in branch geography progressively materialized following a series of deregulations in the 1980s, early 1990s, and ultimately, the 1994 Riegle-Neal Interstate Banking and Branching Efficiency (IBBE) Act. The latter permitted interstate branching in almost all states as of June 1997.

Cases can be made for and against the greater distances between the locations of a bank's activities that follow from the geographical expansion of its branch network. On one hand, through the technology banks have access to, they are equipped to manage the issues associated with these greater distances to support head office objectives in the new markets. This is because such technology can reduce the agency cost of a multibank holding company when there are greater distances between the parent bank and its affiliates (Berger & DeYoung, 2006). This reduced agency cost follows from the technology enabling (i) senior managers at a bank's headquarters to more effectively monitor and communicate with staff at distant branches/subsidiaries, which recently played a key role during the work-from-home COVID restrictions; (ii) more efficient interactions with customers over longer distances; (iii) greater use of quantitative methods from applied finance that aid lending to borrowers without geographical proximity; and (iv) the use of financial engineering products that allow banks (independently of the distance to the counterparty) to unbundle, repackage, or hedge risks at low cost. In contrast, it has been noted that the technology that has enabled banks to improve loan decisions and monitoring over greater distances (i.e., (iii) above) has reduced the need for banks to expand geographically to grow (e.g., DeYoung et al., 2004). This reduced need for branches to be near their customers has been reinforced by the rise of internet banking and in more recent years is further evidenced by this rise leading to branch closures.

Related to greater distances between a bank's locations is the extent of the resulting increase in the geographical diversification of its business (e.g., Chu et al., 2020; Goetz et al., 2016, and Levine et al., 2021). There are also compelling reasons for and against the geographical diversification of a bank's activities that follow from the geographical expansion of its branch network. The benefits of the resulting geographical risk diversification can lead to a bank having a better risk-expected return frontier (Berger & DeYoung, 2006), which would enable the bank to earn higher average revenues from a higher risk-expected return investment strategy. Geographical diversification also enables banks to benefit from scale and scope economies, reduced costs, synergy gains, and the improvement in corporate governance stimulated by the increase in the number of potential corporate acquirers (Deng & Elyasiani, 2008). Conversely, geographical diversification can be detrimental to a bank because of the learning costs that are involved. Also, although technology can reduce the agency cost associated with greater distances between a bank's activities, geographical diversification may lead to an increase in the agency cost from other sources, such as the more complex organizational structure and the specifics of regionally differentiated product packages.

Accordingly, there are two main approaches in the literature to account for the geography of a bank's activities. The first uses a variable that measures the (weighted) distance between the locations of a bank's activities, while the second involves constructing a variable that measures the geographical diversification of a bank's deposits or branches. Each variable is then used as a determinant in a model that is estimated in a standard way using non-spatial methods that assume the observations of the outcome variable for a cross-section of banks are independent across space. In our empirical analysis this independence assumption does not hold, and in many banking applications where this assumption is made it may well be invalid. To motivate our use of methods from spatial data science to account for the spatial dependence of the banks' observations of the outcome variable, in the next section we review the details of these two main approaches to account for the geography of a bank's activities. It is very important that this

spatial dependence is not ignored in banking modeling, otherwise (i) valuable further spatial insights may be overlooked; (ii) the statistical inference will be invalid; and (iii) in certain spatial settings, such as our empirical case, parameter estimates (and thus the fitted quantitative relationships between variables) may be biased. Such spatial methods have been developed to account for the spatial dependence of the cross-sectional observations of the outcome variable and have been widely applied to regions, states, and cities that are geographically linked, for example, by a common border or their relative close proximity. Along the same lines, we propose that these methods are well suited to account for the spatial dependence between banks' observations of various outcome variables.

We provide an accessible presentation of the approach to apply methods from dynamic spatial data science to banks. This approach is presented in terms of the dynamic spatial cost function we estimate, where key explanatory variables in our model are the contemporaneous spatial lag of the dependent variable (i.e., the spatial autoregressive, SAR, variable) and its time lag. One motivation for the inclusion of this time lag is the parallels with time lags in time series analysis and non-spatial dynamic panel data econometrics. That is, there may be some time persistence in the cost spillovers between neighboring banks, and so it may take some time for some of these spillovers to occur. As we use quarterly (rather than annual) data, we are in theory more likely to observe persistence in cost spillovers in the form of a significant time lag of the SAR variable. This is because over a shorter time frame the past is more likely to influence the present. In the spatial literature positive spatial dependence is far more common, so it is more likely that we will find that the SAR variable and its time lag will have positive impacts, which would be consistent with neighboring banks' costs being impacted by common economic phenomena, such as industry-wide regulatory policies, market growth, and headline changes in city, state, and regional economies. Although less likely, we may observe that the SAR variable (and/or its time lag) has a negative impact, where in the spatial literature this is attributed to the effects of competition (e.g.,Boarnet & Glazer, 2002; Garrett & Marsh, 2002). The absolute magnitudes of the coefficients on the SAR variable and its time lag will indicate whether there is more evidence of contemporaneous or dynamic SAR dependence in the data.

Our approach is also general and besides cost can also be applied to other banking variables where the geography of activities is regarded as an important determinant (see Section 2 for other such variables). In the presentation of our empirical approach, the discussion focuses on how the type of spatial model we employ yields contemporaneous and dynamic interbank spillovers (i.e., spill-ins and spill-outs) that measure the impacts of the geography of operations. Specifically, we draw on two appealing features of our model. First, the contemporaneous and dynamic spill-in and spill-out elasticities measure how the impacts of the geography of activities are manifested within the impacts of bank variables, such as in our application the impacts of bank outputs. Second, these spill-in and spill-out elasticities can be used to calculate informative post-estimation measures, such as the contemporaneous and dynamic returns to scale spill-ins and spill-outs we compute.

Rather than following either of the two main approaches to account for the geography of a bank's activities and making the invalid assumption that the cross-sectional cost observations of the banks are independent, we build on the growing number of technical efficiency and productivity studies that use methods from spatial data science. A further reason why we do not follow these two approaches is because they yield own (weighted) distance and own geographical diversification effects (see Section 2 for more details on this), while our focus is on the wider industry impacts of banks' geographical operations. That is, using a modeling approach from spatial data science, we draw on the geographical interconnectedness of banks due to, for example, their common loan and deposit markets, and are thus interested in how the geography of a bank's operations impacts other banks and vice versa. In particular, we build on the study of returns to scale in banking by Glass, Kenjegaliev, and Kenjegalieva (2020) in the following respects. We advance the methodology on spatial returns to scale by moving from the exclusively static (i.e., contemporaneous) spatial setting to one that considers both static and time dynamic spatial relationships. This is because overlooking a dynamic spatial relationship (i.e., some persistence of a spatial effect over time) can impact the estimate of a contemporaneous spatial relationship.

We also make two empirical contributions. (i) To analyze the impact of the first portion of the COVID pandemic on spatial returns to scale for large- and medium-sized U.S. banks, instead of annual data, we use recent quarterly

data (1998:Q1 − 2020:Q4). A theoretical feature of our spatial translog model is that it yields heterogeneous spatial elasticities outside the sample mean for each bank-quarter. We exploit this feature to obtain spatial returns to scale for each bank-quarter, which then enables us to compare these returns during the first portion of the pandemic with other periods in our sample (e.g., the 2008 financial crisis). From the perspective of bank regulators and different bank sizes, we are able to comment on whether the spatial returns to scale support different measures during the crisis and the pandemic.[1] (ii) We exploit the richness of the spatial time dynamics by suggesting which of the banks in the top quintile of the size distribution are the top-ranked initiators (recipients) of the biggest returns to scale spill-outs (spill-ins) in current and future periods. This can be related to the stress testing by U.S. bank regulators that will continue to cover the global systemically important banks (G-SIBs) (Financial Stability Board, 2019) and the banks included in the *Comprehensive Capital Analysis and Review* (CCAR) (Federal Reserve Board, 2019). The CCAR banks are a group where a constituent bank has the capability to influence the domestic U.S. banking industry. The findings from (ii) enable us to suggest whether regulators may consider extending the stress testing to include other large banks that we find are prominent in bank cost interdependence. Further in this opening section we summarize the specific findings from (i) and (ii).

We adapt the methods for the own contemporaneous ray-scale economies (*RSE*) and expansion-path scale economies (*EPSE*) (Berger et al., 1987; Wheelock & Wilson, 2001, 2012, 2018) to obtain the corresponding contemporaneous and dynamic spatial measures. This involves transforming our fitted dynamic spatial model to obtain five contemporaneous and dynamic elasticities: direct, two indirect (spill-in and spill-out), and hence two total (LeSage & Pace, 2009). We collectively refer to these elasticities and the corresponding *RSE* and *EPSE* as contemporaneous and dynamic *spatial* elasticities and returns to scale, as they are partially or entirely made up of spill-ins or spill-outs. The contemporaneous and dynamic direct elasticities are akin to own impacts as they measure the effect of a change in a bank's own independent variable in the current period on its costs in current and future periods. Also associated with the change in the bank's own independent variable are contemporaneous and dynamic indirect elasticities that measure the cost spill-in (spill-out) to (from) the bank in current and future periods. Summing the direct and indirect spill-in / spill-out impacts yields two total elasticities. We then use these contemporaneous and dynamic spatial elasticities to construct a series of corresponding cost models. From this series of models we calculate a series of contemporaneous and dynamic spatial *RSE* and *EPSE* measures.

To fix ideas, the only SAR dependence that Glass, Kenjegaliev, and Kenjegalieva (2020) model is contemporaneous. By including a SAR variable, they were able to compute the direct, indirect, and total impacts of a change in a bank's own independent variable in the current period on the bank's above corresponding contemporaneous costs. The principal motivation for extending their approach to also include the time lag of the SAR variable in our model is because it allows us to go a step further and compute the direct, indirect, and total impacts of a change in a bank's own independent variable in the current period on the bank's above corresponding costs in future periods. Using these dynamic impacts we address the important issue of how many future periods it takes for these impacts to die out (see Table 3 in Section 5).

The own cost-oriented *EPSE* measure can be viewed as preferable to the corresponding *RSE* as the former allows for the possibility that a bank lies away from a radial ray in the output space. For the same reason, we have a preference for direct, indirect, and total *EPSE* over the corresponding *RSE*, where total returns to scale incorporate direct and indirect spill-in / spill-out returns. Of these three cases (direct, indirect, and total), we have the strongest preference for indirect *EPSE* over indirect *RSE*, as this is where there is the greatest difference between our *EPSE* and *RSE* findings. We therefore conclude that adapting own *RSE* and *EPSE* to the spatial setting strengthens the case for the *EPSE* method over the *RSE* approach. Based on this stronger preference for *EPSE* over *RSE*, we concentrate on the former. Relatedly, an important motivation for including the time lag of the SAR variable in our model is that it allows us to show that the implications of *EPSE* for the optimal size of a bank, which, in turn, will impact its

[1]We thank an anonymous reviewer for suggesting that for different bank size categories we compare the spatial returns to scale results during the crisis with those for the pandemic.

competitiveness, can be more complex than in the static spatial and non-spatial cases. This is because the contemporaneous and dynamic total elasticities for a variable will likely differ and may well yield contemporaneous and dynamic total *EPSE* that are odds with one another. In this situation, if a bank acts on its contemporaneous total *EPSE* by (not) changing its size in the current period, this can mean suboptimal dynamic returns in future periods, which, in turn, will impact a bank's competitiveness over time. In such a situation, we suggest that a bank optimizes its contemporaneous and dynamic total returns over the time frame of its future plans, which would involve some returns being suboptimal for particular periods within this time frame.

Two of the main findings from our empirical analysis are as follows. First, we find that the most noticeable impacts of the 2020 portion of the COVID pandemic are on banks' indirect spill-in and spill-out *EPSE*, while there was relatively little impact on their direct-own *EPSE*. This finding is in line with the pandemic affecting the industry and not just individual banks. We also observe a clear difference between the impacts of the pandemic on the two contemporaneous indirect *EPSE* for smaller and larger banks (quintiles 1 and 5 of the bank size distribution, respectively). For quintile 1 the pandemic led to these two *EPSE* measures declining further below 1, while for quintile 5 they increased further above 1. From the perspective of the cost spill-in and spill-out interactions of a quintile 1 (quintile 5) bank with the other banks in the sample, this suggests that, on average, the pandemic led to quintile 1 (quintile 5) banks becoming suboptimally smaller (larger). This is consistent with large banks providing most of the required funding to firms in the U.S. when they turned to the banks for the liquidity provision using preexisting lines of credit during the pandemic (Li et al., 2020).

Second, we find that a number of the quintile 5 banks with a high-ranking contemporaneous indirect spill-in and/or spill-out *EPSE* have geographically concentrated branches and/or specialize in particular activities (see the empirical analysis for details of these banks). These high-ranking indirect *EPSE* may be because the geographical and operational focus of such banks is associated with higher-quality service spillovers leading to relatively high cost spillovers. Ongoing stress testing of U.S. banks by regulators will continue to cover G-SIBs and CCAR banks. We, however, find that a number of banks with a high-ranking contemporaneous indirect spill-in and/or spill-out *EPSE* are not G-SIBs or CCAR banks. In terms of the policy implications of these results, regulators may consider extending the stress testing to include the other large banks that we find are prominent in bank cost interdependence.

The remainder of this paper is structured as follows. Section 2 reviews approaches to account for the effect of the geography of a bank's activities. Section 3 has two parts. In the first part we set out the dynamic spatial translog cost model we use. In the second, we discuss how this model is transformed to obtain sets of contemporaneous and dynamic spatial elasticities, which we then use to construct a series of corresponding models. In Section 4, we set out how this series of models is used to calculate a series of contemporaneous and dynamic spatial returns to scale measures. In Section 5, we present the empirical analysis of medium-sized and large U.S. banks. Section 6 concludes with a summary that provides some insights for banks and regulators.

2 | APPROACHES TO ACCOUNT FOR THE GEOGRAPHY OF BANKING ACTIVITIES

We noted in the previous section that there are two main approaches to account for the geography of banks' activities. In this section we (i) review the use of these approaches in selected studies; and (ii) explain how our use of methods from spatial data science to account for the spatial dependence between banks with overlapping branch geography builds on these approaches.

The first of the two main approaches uses a variable that measures the (weighted) distance between the locations of a bank's activities. This variable is used as a determinant in various empirical models to analyze a range of banking research questions. To some extent the earlier of two studies by Berger and DeYoung (2001, 2006) resembles the approach we use here as they estimate a cost model. These two studies analyze the effect of the distances between a U.S. parent bank and its domestic affiliate banks on the (cost and/or alternative profit) efficiencies of the latter

(specifically, the later study uses the affiliates' alternative profit efficiency rankings). To ascertain how distance affects the economic relationship between the parent bank and its affiliates, the distance variable is interacted with the efficiency of the parent. Among other things, both studies analyze whether a relatively efficiently managed parent bank can impart its superior managerial skills and practices on its distant affiliates. This would reduce the agency cost associated with the distance between the parent and its affiliate and is consistent with the latter having a higher efficiency. If the parent encounters problems monitoring junior managers at a distant affiliate bank, the agency cost will be higher, leading to the affiliate having lower efficiency.[2] The earlier study finds, first, that an affiliate's efficiency tends to be higher if it is located in a state and region that is relatively near to its parent and, second, that an affiliate's efficiency tends to decrease the further it is located from its parent. That said, this distance effect on an affiliate's efficiency tends to be modest. This suggests that a relatively efficient banking organization is not associated with a particular geographical scope. A further implication from their results is that relatively efficient parent banks can overcome the negative impact of distance and impart their superior managerial skills and practices on their affiliates. The later study concludes that such findings are consistent with banks employing technological advances in the industry.

Another variable that Degryse and Ongena (2005) find is affected by the geographical locations of banks' activities is the interest rates on loans. Over the period of their analysis, they use information on all the loan contracts between small firms and a large Belgian bank to estimate the effects on loan rates of the distances between the borrowing firm and both the lending bank and the reference competing bank. They report evidence of spatial price discrimination in bank lending, as they find that the loan rate is negatively associated with the distance between the lender and the borrower, and positively associated with the distance between the borrower and the reference competing bank. In contrast, rather than including a pure distance determinant, Deng and Elyasiani (2008) account for the economic impacts of the geography of the activities of U.S. bank holding companies (BHCs) on their value and total risk (i.e., the composite of company specific risk and the systemic risk the company poses). To do so, in any one model, they include one of two weighted average distances as a determinant. In particular, the average distance between the headquarters of a BHC and its branches (subsidiaries) is weighted by the branch deposit share (subsidiary total asset share). On average, they find that greater distance between a BHC's headquarters and its branches is associated with a reduction in the value of the company and an increase in its total risk.

Studies such as those considered above that include the distance between the locations of a bank's activities as a determinant report estimates of the own distance effect. We build on this approach in the following three respects by drawing on the growing number of studies that use methods from spatial data science to analyze returns to scale (Glass, Kenjegaliev, & Kenjegalieva, 2020); technical efficiency (e.g., Algeri et al., 2022; Glass, Kenjegalieva, & Weyman-Jones, 2020; Horrace et al., 2019; Orea & Álvarez, 2019; Tsionas & Michaelides, 2016); and productivity (Glass et al., 2013; Glass & Kenjegalieva, 2019; Glass, Kenjegalieva, & Douch, 2020).[3] All these studies have a static spatial framework and hence focus exclusively on contemporaneous spatial relationships. Our study may therefore initiate dynamic spatial technical efficiency and productivity studies that begin to address this imbalance in the literature. First, although we also report the own effects of a bank's explanatory variables, our main contribution is to report the spillover effects associated with changes in these variables. These interbank spill-ins and spill-outs materialize because of the spatial dependence in the data for banks with overlapping branch geography. Second, resembling to some extent how an interaction term is used in the banking literature to account for the effect of distance on an economic relationship, we account for the effect of overlapping branch geography on economic spillovers using the degree of this overlap to weight a neighboring bank's economic variable. Third, whereas in any one empirical model only a single economic measure (branch deposit share or subsidiary total asset share) has been used to weight the distance determinant, we recognize that a larger number of economic relationships in a model may be impacted by the overlapping branch geography. Therefore, in our model we include a full range of neighboring banks' weighted economic variables as determinants.

[2]Conversely, if the parent is relatively inefficiently managed, there may be a reduced agency cost (or even an agency benefit) associated with the distance between the parent and its affiliate. This is consistent with a smaller mark down of the affiliate's efficiency (or even its efficiency being pushed up).
[3]Other spatial data science studies of technical efficiency are Druska and Horrace (2004); Glass, Kenjegalieva and Sickles (2014, 2016); Gude et al. (2018); Orea et al. (2018); Jin and Lee (2020); and Kutlu et al. (2020).

The second main approach to account for the effect of branch geography involves using a measure of the geographical diversification (concentration) of a bank's activities. A bank's geographical diversification has been accounted for using variables that measure its number of branches and the number of states it operates in (e.g., Aguirregabiria et al., 2016, and ; Zamore et al., 2019), or via a dummy variable that indicates whether it has engaged in such diversification (e.g., Goetz et al., 2013). Moreover, to measure a bank's geographical concentration, Degryse and Ongena (2005) use the banks' number of branches in a postal zone to compute each bank's Herfindahl–Hirschman index (HHI). They then use this measure of a bank's market share in a postal zone to analyze how it impacts the interest rate the bank charges on a loan. Their HHI measure, however, does not account for the size of the branch, for example, the branch deposit level. This is presumably for data availability reasons as their rich data are on individual loan contracts of a large Belgian bank and do not cover deposits. Using the available data for U.S. banks, a widely used measure of the concentration of a bank's activities is the HHI of its branch level deposits across its geographical markets, where these markets are defined as metropolitan statistical areas (MSAs) or non-MSA counties. Berger and DeYoung (2001, 2006) use this measure to analyze how it impacts a bank's efficiency (and efficiency ranking), while Hirtle (2007) estimates how a weighted measure of this HHI affects various U.S. bank performance measures (average deposits per branch; average small business loans per branch; and bank profitability, namely, the return on equity and the risk-adjusted market return).[4] One minus the above branch deposit-based HHI is a widely used measure of the geographical diversification of a bank's activities. This type of measure has been used to analyze the impact of such diversification on various bank outcome variables, including total risk (Deng & Elyasiani, 2008); market value (Deng & Elyasiani, 2008; Goetz et al., 2013); loan quality measures (Goetz et al., 2016); the systemic risk the company poses (Chu et al., 2020); and total funding costs (Levine et al., 2021).

On average, deposit levels at branches in the same geographical market will likely be spatially dependent on one another, and so this spatial correlation will be inherent in the data. The HHI that is used to account for the geographical diversification (concentration) of a bank's activities will therefore capture the spatial dependence of branch deposits. However, including this HHI as a determinant in a model that is estimated in a standard way by assuming spatial independence among a cross-section of banks will yield only the own effect of the HHI. We therefore apply accessible spatial data science methods that take into account the spatial dependence among each cross-section of banks in our panel data. This is in line with this paper reporting not just the own effects of determinants, but focusing particularly on the spillovers associated with these variables. Spatial data science methods are well suited to this task as they were specifically developed to, among other things, estimate spill-ins and spill-outs.

Moreover, although the HHI-based measure of the geographical diversification (concentration) of a bank's activities will capture the spatial dependence of branch deposits, this measure is based on a single bank variable and will therefore reflect only one dimension of the spatial dependencies between banks. By overlooking the other dimensions, there is an omitted variables issue that may lead to biased parameter estimates. To capture the spatial dependencies between banks more fully, we use methods that have been specifically designed for this purpose. In particular, this involves using a full range of neighboring banks' weighted variables as determinants to capture the range of spatial dependencies.

3 | SPATIAL MODELING APPROACH

3.1 | Dynamic spatial cost model

The form of the spatial cost model for panel data that we estimate is set out in Equation (1), where the variables are logged. In the spatial literature, this type of model is referred to as a dynamic spatial Durbin model (SDM).

[4]Specifically, Hirtle weights the standard HHI by a bank's share of the total number of branches in the market.

$$c_{it} = \alpha + TL(y_{it}, p_{it}, t) + STL\left(\sum_{j=1}^{N} w_{ij}(y_{jt}), \sum_{j=1}^{N} w_{ij}(p_{jt})\right) + STL\left(\sum_{j=1}^{N} w_{ij}(y_{jt-1}), \sum_{j=1}^{N} w_{ij}(p_{jt-1})\right) + \delta \sum_{j=1}^{N} w_{ij} c_{jt} + \lambda \sum_{j=1}^{N} w_{ij} c_{jt-1} + \eta_i + \varepsilon_{it}. \quad (1)$$

The data comprise observations for T periods (indexed $t \in 1,...,T$) and N banks (indexed $i,j \in 1,...,N \forall i \neq j$).[5] c_{it} is the total cost observation for the *ith* bank in period t; α is the intercept; y_{it} is the $(1 \times K)$ vector of observations for the outputs (indexed $k \in 1,...,K$); p_{it} is the $(1 \times L)$ vector of observations for the input prices (indexed $l \in 1,...,L$); t is the time counter; $TL(y_{it}, p_{it}, t)$ is the translog function; η_i is a fixed effect to account for unobserved heterogeneity; and ε_{it} is noise.[6] $TL(y_{it}, p_{it}, t)$ therefore denotes that this function consists of the first-order variables in brackets as well as particular functions of these variables, namely, their squared terms and interactions between the output and input prices. Collectively t and t^2 represent a nonlinear time trend that measures Hicks neutral technical change.[7]

W is the $N \times N$ spatial weights matrix and is made up of the nonnegative weights w_{ij}. W is specified a priori and represents (i) the spatial arrangement of the banks in each cross-section, and (ii) the strength of the spatial interaction among these banks. As W applies to each cross-section in the panel, the $NT \times NT$ spatial weights matrix for the whole panel is $I_T \otimes W$, where I_T is the $T \times T$ identity matrix and $\otimes$ is the Kronecker product.[8] Following the vast majority of the spatial literature, the spatial weights in Equation (1) are exogenous. Given this exogeneity, a geographical measure is frequently used to specify the spatial weights. We therefore use the same approach that involves using a novel measure of the spatial interconnectedness of banks' branch networks. $\sum_{j=1}^{N} w_{ij} c_{jt}$ is the *ith* observation of the contemporaneous SAR variable and we also include a time lag of this variable. This time lag is included to reflect that there may be some time persistence in the cost spillovers between neighboring banks and so it may take some time for such spillovers to occur.

A feature of spatial modeling is that there are limits on the SAR parameter(s): $\{\delta, \lambda\} \in (1/g_{\min}, 1/g_{\max})$, where $g_{\min}$ and $g_{\max}$ are the most negative and positive real characteristic roots of W, respectively. Specifically, W is a normalized spatial weights matrix, where as a result of the normalization we use in our empirical analysis $g_{\max} = 1$. For details of this normalization and the empirical specification of W, and also details of the variables and data that we use for the empirical analysis, see Subsection 5.1.

The process to settle on the form of Equation (1) involved two steps. In the first step we chose between the various different spatial model specifications. Having chosen in the first step the SDM, in the second step we settle on the details of our model. We chose the SDM in the first step for three reasons. First, it is well known that the SDM nests the corresponding spatial error and SAR models, where the latter is Equation (1) with the *STL* function and its time lag omitted (see below for discussion of the *STL* function). The SDM will therefore yield unbiased parameter estimates even if the true data-generating process (DGP) is either of these other two spatial models. Second, in the spatial error model the spillovers relate to the disturbance, whereas the spillovers from a model that contains the SAR term have a business and economic interpretation as they relate to the independent variables. Third, the ratio of the indirect spillover and direct-own elasticities from the SAR model is the same for all independent variables, which is unlikely to be valid in empirical applications and is not the case with the SDM.[9]

[5] As is standard in spatial modeling the panel data are balanced. This is for theoretical statistical reasons, namely, the breakdown of the asymptotic properties of spatial panel data estimators when the panel is unbalanced and the reason for the missing data is not known (Elhorst, 2009). This breakdown occurs because for unbalanced panel data the spatial weights matrix, which we will introduce shortly, is not of fixed dimension.

[6] Berger and Mester (1997) introduced the concept of banking netputs, and we explored including equity as a netput. We ultimately omitted equity, and in doing so follow Koetter et al. (2012), as this yields own returns to scale that are more in line with those reported in the literature.

[7] We specify technical change as Hicks neutral by omitting interactions with t. This is for parsimony as modeling the spatial dependencies involves including quite a large number of regressors.

[8] We acknowledge an anonymous reviewer for suggesting us to provide details of $I_T \otimes W$.

[9] A further spatial model specification involves augmenting the corresponding dynamic spatial error model with a SAR term and its time lag. We do not pursue this model for the aforementioned second and third reasons. Note also that we use the branch geographies of the banks to specify W. We are therefore interested in the spatial correlations at the micro level and so only consider model specifications where these correlations die out across space. This is in contrast to a more macro approach that considers the common spatial correlation across units (e.g., common factor models).

The starting point to settle on the details of Equation (1) in the second step is Wheelock and Wilson's (2012, 2018) static non-spatial studies of returns to scale in U.S. banking. They estimate theoretical functions and, as a result, their models are parsimonious as they omit variables that shift the frontier. We therefore omit c_{t-1} from the regressors because although this variable distinguishes between a non-spatial dynamic panel data model and its static counterpart (which is not the case in our spatial setting as the inclusion of $\sum_{j=1}^{N} w_{ij}c_{jt-1}$ makes our model dynamic), c_{t-1} is not part of the theoretical translog function. For the same reason we omit all other non-spatial variables that are not part of the theoretical translog function. However, by construction, our model includes spatial variables that shift the frontier. We confine these to spatial lags of the variables in *TL* that relate to the output and input prices (denoted *STL*), the spatial lag of the dependent variable and its time lag and, to be consistent with the rationale for the inclusion of this time lag (namely, that it takes some time for spillovers to occur), a time lag of the *STL* variables.[10] All these spatial variables in Equation (1) are local spatial regressors as they only account for spatial interaction between a bank and its first-order neighbors. That said, it is the presence of the spatial lag of the dependent variable and its time lag in Equation (1) that lead to the global spatial interaction when this model is transformed into its reduced form (see the discussion in Subsection 3.2). This global spatial interaction (contemporaneous and dynamic in future in-sample periods) is between a bank and its first-order neighbors, second-order neighbors (i.e., a neighbor of neighbor), etc.

We estimate Equation (1) using the quasi-maximum likelihood approach in Yu et al. (2008). Three further salient features of the estimation procedure are as follows. (i) As is standard, we use the within transformation to eliminate the fixed effects. (ii) Our approach corrects for the biases from the fixed effects in the model, which are akin to the well-known biases in the non-spatial dynamic setting (see Nickell, 1981). (iii) As is standard in spatial modeling, part of the estimation involves the transformation from the error term to the dependent variable. This transformation accounts for the endogeneity of the contemporaneous SAR variable, and also the fact that the error term is not observed (Anselin, 1988; Elhorst, 2009).[11]

3.2 | Sets of elasticities and the related translog functions

Having estimated Equation (1), the estimate of the *TL* function yields own elasticities for the variables at and outside the sample mean. The estimates of the coefficients on the SAR variable and its time lag, and the elasticities at and outside the sample mean from the estimates of *STL* and its time lag, are elasticities that represent local spillovers to a bank from marginal changes in the spatially weighted contemporaneous and dynamic independent variables of its first-order neighbors. These elasticities do not therefore represent the global spillovers to a bank from its higher order (as well as its first-order) neighbors. We draw on an appealing feature of our model to overcome this shortcoming by computing five further elasticities from the spatial literature that are partially/entirely made up of a global spillover, namely, direct, two indirect (spill-out and spill-in) and thus two total elasticities. This first involves indexing the time horizons in our sample $\gamma \in 0,\ldots,\Gamma = T-1$, where the five elasticities measure the contemporaneous and dynamic cost impacts (i.e., in horizon 0 and the remaining future in-sample horizons, respectively) of a marginal change in an *ith* bank's variable in period *t*. As we discuss further below, when an *ith* bank's variable changes in period *t* these elasticities give rise to five different types of cost change. Specifically, we obtain the five elasticities by taking the partial derivative of the reduced form of Equation (1) (i.e., the DGP) with respect to each variable in the

[10]The variables in the *STL* function and its time lag in Equation (1) are in brackets to indicate that, in addition to spatially lagging the variables, we spatially lag functions of the variables (e.g., y_{jt}^2 and y_{jt-1}^2).

[11]As is standard in micro panels, the number of banks is large $(N = 403)$. In non-spatial panel data analysis *T* is often small, and so in a dynamic setting the time lag of the dependent variable is a source of endogeneity. When this is the case for a dynamic spatial model, the time lag of the SAR variable will also be endogenous. In our spatial panel data *T* is not small (92 quarters), and so following Yu et al. (2008) the time lag of the SAR variable is taken to be exogenous.

TL function. Below we provide a non-technical discussion of the direct, indirect, and total elasticities, but for technical details on the calculation of these elasticities, see Debarsy et al. (2012).[12]

A contemporaneous *direct* elasticity measures the impact of a marginal change in an *ith* bank's variable in period *t* on the same bank's cost in horizon 0. This elasticity has two components—the standard own elasticity from Equation (1) and feedback. This feedback is the contemporaneous effect of a change in an *ith* bank's variable that reverberates back to the same bank's dependent variable through its effect on the dependent variables of the other banks in the sample. In the spatial literature this feedback is typically small (e.g., Autant-Bernard & LeSage, 2011).

There are two contemporaneous *indirect* elasticities that measure the bidirectional spillover impacts of a marginal change in an *ith* bank's variable in period *t*: (i) the spill-out from the *ith* bank to the dependent variables of all the other banks in the sample in horizon 0; and (ii) the spill-in to the dependent variable of the *ith* bank in horizon 0 from all the other banks. Having computed (i) and (ii) for each bank, and as suggested by LeSage and Pace (2009), to facilitate interpretation we report an average of (i) or (ii) across all the banks. It follows from the formulas for the two contemporaneous indirect elasticities that averaging (i) or (ii) across all the banks (at the sample mean or at some other point in the sample) yields the same value, that is, symmetric bidirectional contemporaneous indirect elasticities.[13] Because the average indirect (i) and (ii) elasticities across all the banks are equal, we obtain the same average value for the contemporaneous indirect spill-out and spill-in returns to scale.

There are two *total* impacts in horizon 0 of a marginal change in an *ith* bank's variable in period *t*. This is because a contemporaneous total elasticity is the sum of the contemporaneous direct and indirect (i) or (ii) elasticities. The first of the total impacts is the sum of the direct impact on the *ith* bank's cost and the indirect cost spill-out from this bank to all the other banks. The second is the sum of the direct impact and the indirect cost spill-in to the *ith* bank from all the other banks. We therefore obtain the same value for the average total elasticity across all the banks when we use either the average indirect (i) or (ii) elasticity.

However, for any subset of banks the two contemporaneous indirect elasticities will be of different magnitudes. We calculate these asymmetric elasticities for individual banks and also bank size categories. As a result, for these banks, the contemporaneous total elasticities and contemporaneous indirect and total returns to scale are also asymmetric.

Along the same lines, we calculate average measures across all the banks of the corresponding dynamic elasticities (direct, symmetric indirect spill-out and spill-in, and the resulting two symmetric total measures). These dynamic elasticities for individual banks measure the impacts of a marginal change in the *ith* bank's variable in period *t* on the same costs as those in the above contemporaneous impacts but in future in-sample time horizons $(\gamma \in 1,...,\Gamma = T-1)$. These elasticities for individual banks are then used to calculate average dynamic direct, indirect, and total returns to scale for bank size categories. Given the importance of the too-big-to-fail (TBTF) banks for the stability of the banking system, to investigate their spatial interactions (and those of other large banks) we focus on their asymmetric contemporaneous and dynamic indirect returns.

We conduct the statistical inference for the contemporaneous and dynamic direct, indirect, and total parameters using Monte Carlo simulations. This involves drawing 500 Halton sequences of parameter values from the variance–covariance matrix, where each value has a random component drawn from $N(0,1)$.

Using the estimates of the direct, two indirect (spill-out and spill-in) and thus two total parameters for each variable, one can specify five translog equations for the associated costs in horizon γ; namely: the *ith* bank's direct cost $\left(c^{Dir}_{i,\gamma}\right)$; the indirect cost spill-out from the *ith* bank to the other banks $\left(c^{Ind}_{Out,i,\gamma}\right)$; the indirect cost spill-in to the *ith* bank from the other banks $\left(c^{Ind}_{In,i,\gamma}\right)$; and thus the two total cost measures $\left(c^{Tot}_{Out,i,\gamma} \text{ and } c^{Tot}_{In,i,\gamma}\right)$. For the form of the translog equation for $c^{Ind}_{In,i,\gamma}$, see Equation (2). We do not present the forms of the translog equations for $c^{Dir}_{i,\gamma}, c^{Ind}_{Out,i,\gamma}, c^{Tot}_{Out,i,\gamma}$ and $c^{Tot}_{In,i,\gamma}$ as they are similar to Equation (2). This is because the independent variables are the same

[12]We calculate the five elasticities by modifying the approach in Debarsy, Ertur, & LeSage to account for our model omitting c_{t-1} and including time lags of the variables in the *STL* function.

[13]The equality of the averages of (i) and (ii) across all the banks is because the averages of the column and row sums of the off-diagonal elements of a matrix are equal.

as in Equation (2), with only the dependent variable changing and the subscripts and superscripts of the parameters (β and τ) and vectors (κ' and ξ') and matrices (Υ, Ω, and Λ) of parameters matching those of the dependent variable. In the discussion below we relate Equation (1) and its reduced form to the forms of the five translog equations and, as will become clear in Section 4, we use these five equations to compute the direct, indirect, and total returns to scale.

$$\begin{aligned} c^{Ind}_{In,i,\gamma} &= \beta^{Ind}_{In,i,\gamma} t + \frac{1}{2}\tau^{Ind}_{In,i,\gamma} t^2 + \kappa^{Ind'}_{In,i,\gamma} y_{it} + \xi^{Ind'}_{In,i,\gamma} p_{it} + \frac{1}{2} y'_{it} \Upsilon^{Ind}_{In,i,\gamma} y_{it} \\ &\quad + \frac{1}{2} p'_{it} \Omega^{Ind}_{In,i,\gamma} p_{it} + y'_{it} \Lambda^{Ind}_{In,i,\gamma} p_{it}. \end{aligned} \tag{2}$$

In Equation (2), we attach a subscript γ to the parameters and the dependent variable to indicate that a parameter represents the impact in horizon γ of a marginal change in an *ith* bank's independent variable in period t. If t and γ correspond to the same period, which will only be the case when $\gamma = 0$, then the parameters in Equation (2) are contemporaneous. Otherwise, the parameters in Equation (2) are dynamic, which indicates that a change in an *ith* bank's independent variable in t will impact the dependent variable γ (in-sample) periods ahead. Note that a subscript i is attached to the parameters in Equation (2) to denote that the parameters are for an individual bank. Alternatively, and as we have previously noted, these parameters can be averages across all the banks.

There are, however, important differences between the five translog equations and Equation (1). First, whereas the cost variable in Equation (1) is observed, this is not the case for the dependent variable in each of the five translog equations and, as a result, these five equations are not regressions and do not have a disturbance term. We can though use the five translog equations to compute these dependent variables, where $c^{Tot}_{Out,i,\gamma} = c^{Dir}_{i,\gamma} + c^{Ind}_{Out,i,\gamma}$ and $c^{Tot}_{In,i,\gamma} = c^{Dir}_{i,\gamma} + c^{Ind}_{In,i,\gamma}$. Second, Equation (1) includes the SAR variable and its time lag, and also the contemporaneous and dynamic spatially lagged variables in the two *STL* functions. However, in the five translog equations the impacts of these variables are accounted for within the calculations of the direct, indirect, and total parameters.

4 | MEASURING SPATIAL SCALE ECONOMIES

From the relevant translog cost function for horizon γ (e.g., Equation (2) for the indirect cost spill-in to the *ith* bank from all the other J banks in the sample), we calculate the five (direct, indirect spill-in and spill-out, and thus two total) spatial measures of both *RSE* and *EPSE*. Note that these measures are contemporaneous for time horizon $\gamma = 0$ and dynamic for $\gamma = 1, \ldots, \Gamma$. For brevity we present the methods for the five spatial *RSE* and *EPSE* in the context of one of the five cases, namely, indirect spill-in *RSE* and *EPSE* for horizon γ using the corresponding translog cost function. We can do this because this method can easily be adapted to calculate each of the other spatial *RSE* and *EPSE* by replacing the notation for indirect spill-in ($^{Ind}_{In,i,\gamma}$) with that for direct ($^{Dir}_{i,\gamma}$), indirect spill-out ($^{Ind}_{Out,i,\gamma}$), or each of the two total cases ($^{Tot}_{In,i,\gamma}$ and $^{Tot}_{Out,i,\gamma}$).

To set out the interpretations of the five spatial *RSE* and *EPSE*, consider a change in (y_{it1}, y_{it2})—a two-dimensional bundle of outputs for the *ith* bank in period t. (i)$RSE^{Dir}_{i,\gamma}$ and $EPSE^{Dir}_{i,\gamma}$ measure the impact of a change in (y_{it1}, y_{it2}) on the *ith* bank's direct cost in horizon γ from the corresponding translog function (i.e., the bank's own cost plus any feedback to this cost). (ii) $RSE^{Ind}_{Out,i,\gamma}$ and $EPSE^{Ind}_{Out,i,\gamma}$ measure the impact in γ of a change in (y_{it1}, y_{it2}) on the costs of all the other J banks, or, in other words, the impact on the indirect cost spill-out from the *ith* bank. (iii) $RSE^{Ind}_{In,i,\gamma}$ and $EPSE^{Ind}_{In,i,\gamma}$ measure the impact in γ of a change in (y_{it1}, y_{it2}) on the cost of the *ith* bank, where this impact is due to an indirect cost spill-in to the *ith* bank from all the other banks. (iv) $RSE^{Tot}_{Out,i,\gamma}$ and $EPSE^{Tot}_{Out,i,\gamma}$ measure the impact in γ of a change in (y_{it1}, y_{it2}) on the first of the two measures of the *ith* bank's total cost. The two drivers of this cost impact are the direct impact from (i) and the indirect spill-out impact from (ii). (v) $RSE^{Tot}_{In,i,\gamma}$ and $EPSE^{Tot}_{In,i,\gamma}$ measure the impact in γ of a change in (y_{it1}, y_{it2}) on the second measure of the *ith* bank's total cost. The two drivers of this impact are the direct impact and the indirect spill-in impact from (iii).

There are two important differences to highlight between the *RSE* and *EPSE* in (i)–(v). First, these *RSE* measures assume that the full range of output levels of the *ith* bank lies on a radial ray, whereas the *EPSE* measures relax this assumption by considering changes in the *ith* bank's output levels along its output expansion-path. Second, these *RSE* measures relate to an equiproportional change in y_{it1} and y_{it2}, while the *EPSE* measures are concerned with incremental changes in these outputs.

Turning now to the formal presentation of $RSE^{Ind}_{In,i,\gamma}$ and $EPSE^{Ind}_{In,i,\gamma}$.

Spatial RSE

We compute $RSE^{Ind}_{In,i,\gamma}$ as follows:

$$RSE^{Ind}_{In,i,\gamma} = \sum\nolimits_{k=1}^{K} \frac{\partial c^{Ind}_{In,i,\gamma}(t, p_{it}, y_{it})}{\partial y_{kit}}. \tag{3}$$

The elasticity $\partial c^{Ind}_{In,i,\gamma}(t, p_{it}, y_{it})/\partial y_{kit}$ is the first-order derivative of the translog function for $c^{Ind}_{In,i,\gamma}$ (Equation (2)) with respect to the *kth* output of the *ith* bank at *t*.

We know from production theory that own cost *RSE* is positive and $<, =$ or > 1, corresponding to increasing, constant, or decreasing returns to scale, respectively. However, there is no such theory that posits whether any of the spatial *RSE* should be positive or even negative. If an estimate of any of these five *RSE* is positive, then the above returns to scale classification for the own case applies. Alternatively, if any of these estimates is negative, namely, $<, =$ or > -1, this corresponds to increasing, constant, or decreasing returns to scale. The classification of the five spatial *RSE* need not be the same.

The five spatial *RSE* we consider are partially/entirely made up of a *RSE* spill-in/spill-out. Of these measures, $RSE^{Ind}_{In,i,\gamma}$ and $RSE^{Ind}_{Out,i,\gamma}$ are made up entirely of a *RSE* spill-in and spill-out. As a result, for these two measures, the returns to scale classification is based only on the sign and magnitude of the *RSE* spill-in/spill-out. $RSE^{Dir}_{i,\gamma}$ is made up of own *RSE* and feedback *RSE*. If the feedback *RSE* is negative, the issue is whether they more than offset the positive own *RSE*, leading to negative $RSE^{Dir}_{i,\gamma}$. As the feedback parameter estimates in the empirical spatial literature are typically small, we would expect the feedback *RSE* to be small. Consequently, if the feedback *RSE* is negative, we would expect it to be more than offset by the positive and non-negligible own *RSE*, leading to positive and non-negligible $RSE^{Dir}_{i,\gamma}$.

As a result of the above method to calculate the five spatial *RSE*, and the two total elasticities for a variable being the sum of its direct and indirect spill-in/spill-out elasticities, $RSE^{Tot}_{In,i,\gamma} = RSE^{Dir}_{i,\gamma} + RSE^{Ind}_{In,i,\gamma}$ and $RSE^{Tot}_{Out,i,\gamma} = RSE^{Dir}_{i,\gamma} + RSE^{Ind}_{Out,i,\gamma}$. We expect $RSE^{Dir}_{i,\gamma}$ to be positive and non-negligible, and if $RSE^{Ind}_{In,i,\gamma}$ and $RSE^{Ind}_{Out,i,\gamma}$ are negative, the issue is whether they more than offset $RSE^{Dir}_{i,\gamma}$, leading to negative $RSE^{Tot}_{In,i,\gamma}$ and $RSE^{Tot}_{Out,i,\gamma}$.

There is a clear relationship between the SAR parameters and $RSE^{Ind}_{In,i,\gamma}$ (and $RSE^{Ind}_{Out,i,\gamma}$). As a result, the reason for the sign of a SAR parameter is also the business/economic explanation for the sign of $RSE^{Ind}_{In,i,\gamma}$ ($RSE^{Ind}_{Out,i,\gamma}$). The sign of a SAR parameter is determined by whether a bank's observations for the dependent variable are positively/negatively associated with its neighboring banks' (spatially weighted) observations of the same variable. In the empirical spatial literature more generally, said negative association is attributed to the effects of competition (Boarnet & Glazer, 2002; Garrett & Marsh, 2002). In the context of our empirical analysis, such negative SAR dependence would point to spatial cost competition between banks, whereby a change in the cost competitiveness between banks would result in a decrease (increase) in a bank's cost relative to the spatially weighted costs of its neighboring banks. Conversely, said positive association in the spatial literature is attributed to common business/economic phenomena across neighboring firms, such as market growth and headline changes in economies (city, state, regional, and national). In the context of our empirical analysis, a bank's cost can be positively associated with the (spatially weighted) costs of its neighboring banks due to, for example, the common cost implications for banks of regulatory and monetary policies.

To conduct the statistical inference, the Halton parameter sequences (see Subsection 3.2) are used to compute 500 estimates for horizon γ of the five spatial *RSE*.

Spatial EPSE

Own and spatial *RSE* are convenient measures of returns to scale, and own *RSE* is the most reported measure. However, *RSE* measures may not be the most appropriate because a bank may not lie on a radial ray. To accommodate this situation, Berger et al. (1987) and Wheelock and Wilson (2001, 2012, 2018) estimate own *EPSE*, which relate to an incremental move along a (radial or non-radial) portion of a bank's output expansion-path.[14]

We consider five (contemporaneous and dynamic) spatial *EPSE*, which we calculate by applying the method for the non-spatial *EPSE* to each of the corresponding translog cost functions (e.g., Equation (2) is used to calculate $EPSE^{Ind}_{In,i,\gamma}$). Our presentation of the method for the five spatial *EPSE* is in the context of $EPSE^{Ind}_{In,i,\gamma}$. This first involves considering in the space (t, p_{it}, y_{it}) a further point (t_n, p_{itn}, y_{itn}), where to highlight the difference between the five spatial *RSE* and *EPSE* we use the subscript *n* to indicate that the point lies on a non-radial portion of a bank's output expansion-path. $EPSE^{Ind}_{In,i,\gamma}$ measures the expected change in $c^{Ind}_{In,i,\gamma}$, when in period *t*, a bank moves incrementally along a non-radial portion of its output expansion-path between the points $(t_n, p_{itn}, (1-\psi)y_{itn})$ and $(t_n, p_{itn}, (1+\psi)y_{itn})$.

Using Equation (2) we compute $EPSE^{Ind}_{In,i,\gamma}$ as follows:

$$EPSE^{Ind}_{In,i,\gamma} = \frac{c^{Ind}_{In,i,\gamma}(t_n, p_{itn}, (1-\psi)y_{itn})}{\zeta c^{Ind}_{In,i,\gamma}(t_n, p_{itn}, (1-\psi)y_{itn})} = \frac{c^{Ind}_{In,i,\gamma}(t_n, p_{itn}, (1+\psi)y_{itn})}{\left(\frac{1+\psi}{1-\psi}\right) c^{Ind}_{In,i,\gamma}(t_n, p_{itn}, (1-\psi)y_{itn})}. \tag{4}$$

As the relative proportions of the *ith* bank's *K* outputs remain constant, then $\zeta(1-\psi)y_{itn} = (1+\psi)y_{itn}$. Therefore, $\zeta = (1+\psi)/(1-\psi)$ and the latter part of Equation (4) follows.

Following Wheelock and Wilson's (2012) analysis of the own returns to scale of U.S. banks, we use $\psi = 0.05$ to calculate the five spatial *EPSE* for horizon γ. Specifically, we calculate these spatial *EPSE* measures for a movement between $\pm\psi$ (or, in other words, 95% and 105%) of the mean output vector of the full sample or the relevant subsample.

There is no production theory that posits whether any of the spatial *EPSE* measures should be positive or even negative. The returns to scale classification of the five spatial *EPSE* for horizon γ is therefore the same as the above classification of the corresponding spatial *RSE*. Therefore, a positive spatial *EPSE* estimate $<, =$ or $>$ 1 corresponds to increasing, constant, or decreasing returns to scale. Conversely, a negative spatial *EPSE* estimate $<, =$ or > -1 points to increasing, constant, or decreasing returns to scale. The classification of estimates of the five spatial *EPSE* need not be the same. A further similarity between the spatial *RSE* and *EPSE* is that the two business/economic explanations we provided above for a positive or negative spatial *RSE* are also the explanations for the sign of a spatial *EPSE* measure. There is, however, an important difference between the spatial *RSE*, $EPSE^{Tot}_{In,i,\gamma}$ and $EPSE^{Tot}_{Out,i,\gamma}$ are not the sum of the direct and indirect spill-in/spill-out *EPSE*. This is because the spatial *EPSE* are ratios with different denominators.

Statistical inference for the spatial *EPSE* for horizon γ involves following the above approach for the spatial *RSE* by computing 500 estimates.

[14]Non-spatial and spatial *EPSE* can also be thought of in terms of an incremental move along a portion of a bank's input expansion-path, for example, from a (non-spatial or appropriate spatial) revenue function.

5 | EMPIRICAL ANALYSIS OF MEDIUM-SIZED AND LARGE U.S. BANKS

5.1 | Data and the spatial weights matrix

We estimate Equation (1) using quarterly balanced panel data for 403 medium-sized and large U.S. banks over the period 1998:Q1 − 2020:Q4.[15] We focus on medium-sized and large banks because their branch networks are sufficiently large and so there is a sufficient overlap between their networks; that is, there is no lack of interconnectedness between the banks in the spatial weights matrix (see further in this subsection for details of our a priori construction of this matrix). Banks are included in our sample based on the total assets thresholds of the U.S. bank size categories in Berger and Roman (2017). They convert their monetary variables into 2012:Q4 U.S. dollars using the GDP deflator, and we do the same for banks' total assets. We include banks based on their real total assets in the final period of our sample. In this period, we classify a bank with real total assets between $1 billion and $3 billion as medium-sized and greater than $3 billion as large. Based on these classifications both size categories are well represented in our sample (188 medium-sized banks and 215 large). In terms of the charter types of the banks, the sample comprises 97 commercial federal charter Fed member banks, 162 commercial state charter Fed nonmember banks, 86 state charter commercial or savings Fed member banks, 41 savings state charter banks, and 17 savings associations. The vast majority of these banks (338) specialize in commercial lending. Moreover, our study period is interesting because it includes very different bank operating environments, for example, the financial crisis, sufficiently long pre- and post-crisis periods, and the COVID-induced economic downturn in 2020.

The data for the variables are from the Call Reports and were obtained from the Federal Deposit Insurance Corporation (FDIC). This data are at the bank level, and the classification of variables as output and input prices is based on the intermediation approach (e.g., Koetter et al., 2012).[16] In Table 1, we provide a description of the variables and summary statistics. Summarizing, there are three outputs and three input prices. The outputs measure the levels of net loans and leases (y_1), securities (y_2), and non-interest income (y_3). Non-interest income captures nontraditional banking operations, such as the financial product innovations, derivatives, and securitization that occur off the balance sheet. According to Stiroh (2004) and Laeven and Levine (2007), these off-balance sheet activities have a substantial influence on banking performance, while in a bank cost function Clark and Siems (2002) highlight both the statistical and economic importance of the inclusion of both non-interest income as an output and traditional balance sheet outputs.[17]

The input prices are the prices of fixed assets and premises (p_1), labor (p_2), and deposits (p_3). c denotes total operating cost, which we measure as the sum of the expenditures on the three inputs. Following Wheelock and Wilson (2018), we first difference to obtain the data for the expenditure and other flow variables for quarters $2-4$. We then convert c and y_1-y_3 into 2012:Q4 U.S. dollars using the GDP deflator, but not p_1-p_3 because, as we can see from Table 1, they are ratios. We then transform the variables by, in turn, taking logs, mean adjusting and using p_1 as the normalizing factor for c and the other input prices. By mean adjusting the data, we can interpret the contemporaneous and dynamic direct, two indirect, and two total parameters on a first-order variable as elasticities at the sample mean. The heterogeneity of the banks and thus dispersion in the data is evident from Table 1 as the standard deviation of all but two variables (p_2 and p_3) is relatively large compared to the mean. Recall therefore that bank fixed effects are used to account for unobserved heterogeneity.

To recap, we do not follow either of the two main approaches reviewed in Section 2 to account for the geography of banks' operations. Therefore, we do not include a measure of the (weighted) distance between a bank's

[15]The Riegle-Neal IBBE Act of 1994 provided greater opportunities for banks to have overlapping branch networks, leading to greater potential spatial dependence between banks. Our study period starts 6 months after the last 12 states implemented the IBBE Act on 6/1/97 (Dick, 2006). This 6-month lag was to allow a short period of time for out-of-state banks to begin expanding their branch networks in the 12 laggard states. Doing so allows some time for greater spatial dependence between the banks to materialize, that is, significant coefficients on the contemporaneous and dynamic SAR variables.

[16]The intermediation approach is well established and is due to the seminal work by Sealey and Lindley (1977).

[17]We thank an anonymous reviewer for suggesting us to discuss further the inclusion of non-interest income as an output.

TABLE 1 Variable descriptions and summary statistics

Variable description	Model notation	Mean	Std. dev.	Minimum	Maximum
Total operating cost (000s of 2012:Q4 U.S. dollars):	c	108,413	611,068	96	12,079,107
Sum of salaries, interest expenses on deposits and					
expenditure on fixed assets and premises					
Input prices					
Cost of fixed assets and premises: Expenditure on	p_1	0.184	3.396	0.001	253.789
fixed assets and premises divided by their value					
Cost of labor: Salaries divided by the number of	p_2	18.698	7.388	0.732	231.877
full-time equivalent employees					
Cost of deposits: Interest expenses on deposits	p_3	0.004	0.003	9.00×10^{-6}	0.036
divided by deposits					
Outputs					
Net loans and leases (000s of 2012:Q4 U.S. dollars)	y_1	10,334,205	59,849,241	2,524	922,974,016
Securities (000s of 2012:Q4 U.S. dollars)	y_2	3,926,742	24,872,499	233	580,260,352
Non-interest income (000s of 2012:Q4 U.S. dollars)	y_3	92,488	612,964	3	13,280,522

physical operations or the HHI of a bank's branch deposits across its geographical markets as an explanatory variable in a standard non-spatial model. This is because these approaches would only yield own distance and own HHI effects, while our focus is on the wider industry impacts of banks' geographical operations. That is, we draw on the geographical interconnectedness of banks due to, for example, their common loan and deposit markets, and are thus interested in how the geography of a bank's operations impacts other banks and vice versa. Given this focus we use a model from spatial data science that involves the a priori specification of the links between each pair of banks (i.e., the w_{ij}'s that make up W).

Two factors influence our specification of W. First, following the vast majority of the spatial literature the w_{ij}'s in Equation (1) are exogenous and in line with this, we specify W using a measure that reflects the geographical links between banks' operations. Second, we acknowledge that Corrado and Fingleton (2012) recommend that W has some economic foundation. We first rule out a specification of W based on the distances between banks' headquarters as the locations of two banks' headquarters may not be a good indicator of the geographical linkages between their branch networks. We also adopt a cautious approach and do not use the ratios of the ij-th banks' deposits across their geographical markets as spatial weights. This is because these ratios would be measures of the economic distances between neighboring banks and may well therefore be endogenous. Put another way, there are parallels between using such deposit ratios and endogenous spatial weights based on trade flows in a country or regional spatial production function.

We follow, among others, Glass, Kenjegaliev, and Kenjegalieva (2020) and use a specification of W that reflects the degree of the geographical overlap between banks' branch networks, where we use a broad definition of

overlapping networks (see the below details on how we construct W). That is, we adopt a conservative approach and regard banks to have overlapping branch networks if they have branches in the same state. This is because if we adopt the same approach at the county level, we could potentially overlook some linkages between banks. Our specification of W avoids assuming that each bank's neighborhood set includes the same arbitrary number of nearest banks; avoids choosing an arbitrary radius within which banks are regarded as neighbors; is geographical in nature which is consistent with W being exogenous; and the branch geography on which W is based underpins economic linkages between banks in the form of their branch deposits in the same markets.

W is a normalized mean quarterly matrix. In this mean matrix before normalization, which we denote $\tilde{W}$, each element on the main diagonal is set to zero. This is because a bank cannot belong to its set of neighboring banks. To calculate each off-diagonal element of $\tilde{W}$, we index the number of states and Washington DC $v \in 1,...,51$ and use Equation (5).

$$\tilde{w}_{ij} = \frac{\sum_{t=1}^{T} \sum_{v=1}^{51} \frac{\text{Number of } jth \text{ bank branches in state } v \text{ in period } t}{\text{Number of } ith \text{ bank branches in state } v \text{ in period } t}}{T} \forall i \neq j. \tag{5}$$

Relative to the *ith* bank's mean branch network, $\tilde{w}_{ij}$ represents the mean branch network intensity of the *jth* bank. To calculate $\tilde{w}_{ij}$ we use the state locations of banks' branches in the *Summary of Deposits* from the FDIC. The data for the variables are quarterly, rather than the annual data that are also available, so that, among other things, we can have a more high frequency first look at the own and spatial returns to scale over the COVID affected 2020. However, the locations of the banks' branches is annual mid-year information. As this information is at mid-year intervals, and not as with year-end data for the final quarter, we reconcile the quarterly data for the variables with the branch location information by assuming that the latter applies to each quarter in the year. In line with the estimation of the model at the sample mean $\tilde{w}_{ij}$ is an average across the T periods. Finally, to obtain the specification of W we normalize $\tilde{W}$ by dividing throughout by its largest element (i.e., its largest eigenvalue). This normalization is appealing because it does not change the proportional relationship between the spatial weights. Therefore, in the context of our empirical analysis, this normalization retains the information on the relative intensities of the branch networks.[18]

5.2 | Estimated spatial cost model

In Table 2, we present our estimated dynamic spatial cost model (Equation 1). Further in this subsection we analyze the cumulative indirect parameters from the reduced form of this estimated model, which, as noted in Subsection 3.2, measure the global spillovers pertaining to a bank's first-order and higher-order neighbors. The estimates of the coefficients on Wc_t and Wc_{t-1} in Table 2, however, are local elasticities that measure the contemporaneous and dynamic SAR spillovers to a bank from only its first-order neighbors. These two parameter estimates are non-negligible and significant at the 1% level, which supports accounting for contemporaneous and dynamic SAR dependencies in our model. As is common in the spatial literature and thus as we would expect, these two parameter estimates are positive, which is consistent with neighboring banks' costs being impacted by common economic phenomena, such as regulatory and monetary policies, market growth, and headline changes in economies (city, state, regional, and national). We find that there is a non-negligible difference between the coefficient on Wc_{t-1} and its smaller counterpart on Wc_t, which indicates that the largest cost spillovers take some time to occur. This also indicates that there is a relatively high degree of SAR persistence in the cost data. As the coefficients on the SAR variable and its time lag are both positive, from an economic perspective this is consistent with rigidities in bank cost interdependence. As

[18]In the spatial literature the weights matrix is often normalized by its row sums. This is appropriate for binary spatial weights that represent, for example, contiguous geographical areas. As our weights are non-binary, we do not use this normalization. This is because row-normalizing our weights would remove the information on the relative bank branch intensities.

TABLE 2 Estimated dynamic spatial cost model

	Model coeff		Model coeff		Model coeff
Wc_t	0.219***	t	−0.002***	Wy_{1t-1}	−0.316***
Wc_{t-1}	0.30***	t^2	0.000	Wy_{2t-1}	−0.078
y_{1t}	0.56***	Wy_{1t}	−0.149	Wy_{3t-1}	−0.035
y_{2t}	0.166***	Wy_{2t}	0.295***	Wp_{2t-1}	0.042
y_{3t}	0.153***	Wy_{3t}	−0.156***	Wp_{3t-1}	−0.255***
p_{2t}	0.613***	Wp_{2t}	−0.171***	Wy^2_{1t-1}	−0.014
p_{3t}	0.294***	Wp_{3t}	0.125***	Wy^2_{2t-1}	0.001
y^2_{1t}	0.044***	Wy^2_{1t}	0.040	Wy^2_{3t-1}	−0.007
y^2_{2t}	0.016***	Wy^2_{2t}	0.017	$Wy_{1t-1}y_{2t-1}$	0.013
y^2_{3t}	0.035***	Wy^2_{3t}	0.024	$Wy_{1t-1}y_{3t-1}$	0.024
$y_{1t}y_{2t}$	−0.031***	$Wy_{1t}y_{2t}$	−0.062	$Wy_{2t-1}y_{3t-1}$	−0.001
$y_{1t}y_{3t}$	−0.063***	$Wy_{1t}y_{3t}$	0.000	Wp^2_{2t-1}	0.021
$y_{2t}y_{3t}$	0.003***	$Wy_{2t}y_{3t}$	−0.035	Wp^2_{3t-1}	0.035***
p^2_{2t}	0.040***	Wp^2_{2t}	0.007	$Wp_{2t-1}p_{3t-1}$	−0.044***
p^2_{3t}	0.040***	Wp^2_{3t}	−0.033***	$Wy_{1t-1}p_{2t-1}$	0.044
$p_{2t}p_{3t}$	−0.085***	$Wp_{2t}p_{3t}$	0.007	$Wy_{1t-1}p_{3t-1}$	−0.109***
$y_{1t}p_{2t}$	−0.023***	$Wy_{1t}p_{2t}$	0.036	$Wy_{2t-1}p_{2t-1}$	−0.222***
$y_{1t}p_{3t}$	0.033***	$Wy_{1t}p_{3t}$	−0.025	$Wy_{2t-1}p_{3t-1}$	0.159***
$y_{2t}p_{2t}$	−0.008***	$Wy_{2t}p_{2t}$	−0.118***	$Wy_{3t-1}p_{2t-1}$	0.111***
$y_{2t}p_{3t}$	0.016***	$Wy_{2t}p_{3t}$	−0.052*	$Wy_{3t-1}p_{3t-1}$	0.001
$y_{3t}p_{2t}$	0.031***	$Wy_{3t}p_{2t}$	0.087***		
$y_{3t}p_{3t}$	−0.044***	$Wy_{3t}p_{3t}$	0.029**	LL	33234.7

Note: *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

we noted in the opening section, we are in theory more likely to find evidence of such phenomena as we use quarterly (rather than annual) data and over a shorter time frame the past is more likely to influence the present.

In Table 2, the estimated coefficients on the first-order outputs $(y_{1t} - y_{3t})$ and input prices (p_{2t} and p_{3t}) are standard own elasticities at the sample mean and are positive, non-negligible, and significant at the 1% level. As these elasticities are positive, at the sample mean our fitted model satisfies the monotonicity property of a cost function in production theory.

It is also evident from Table 2 that a number of the contemporaneous and dynamic *STL* variables are significant at the 10% level or lower. These findings are supportive of our dynamic SDM specification, as opposed to a dynamic SAR model that would omit the *STL* function and its time lag. A subset of the results for the *STL* function (and its time lag) are for the spatial lags of the first-order output and input prices (and their time lags). We can interpret the coefficients on these variables as local spillover elasticities at the sample mean. Each of these elasticities therefore represents the cost spillover to the hypothetical sample average bank when there is a spatially weighted change in a contemporaneous (dynamic) first-order independent variable of its hypothetical first-order neighbors. Of these variables, the coefficients on Wy_{2t} and Wp_{3t} are positive, non-negligible, and significant, while the coefficients on $Wy_{3t}, Wp_{2t}, Wy_{1t-1}$ and Wp_{3t-1} are negative, non-negligible, and significant. Interestingly, these findings indicate that a decrease in the sample average bank's cost is the spatial competitive effect of a contemporaneous increase in the spatially weighted securities (Wy_{2t}) and cost of deposits (Wp_{3t}) of its first-order neighbors. These findings also

indicate that the sample average bank's cost will increase when there is an increase in the spatially weighted non-interest income (Wy_{3t}) and cost of labor (Wp_{2t}) of its first-order neighbors. As a final point on these findings, we note that the sample average bank's cost will increase in the current period when, in the previous period, there is an increase in the spatially weighted net loans and leases (Wy_{1t-1}) and cost of deposits (Wp_{3t-1}) of its first-order neighbors.

The own and local spillover elasticities from the model in Table 2 represent one of the six sets of elasticities that our model yields. The other five sets account for global spillovers and are sets of direct, two indirect (spill-out and spill-in), and hence two total elasticities from the reduced form of the model. As we noted above, at the sample mean and on average across all the banks outside the sample mean (e.g., for individual quarters), the two indirect and thus two total elasticities for a variable will be equal. For a subset of the banks outside the sample mean, these indirect (total) elasticities will differ in magnitude. In Table 3, for the first-order output and input prices, we present the cumulative direct, indirect, and total elasticities for the current time horizon and first 12 future in-sample horizons $(\gamma = 0,...,12)$. These parameters are elasticities that measure the cumulative impact on the sample average bank's direct, indirect (see Equation (2)) and total cost, with respect to a permanent marginal change in an output (input price) of this bank in the current horizon.

The cumulative direct output and input price elasticities at the sample mean for the contemporaneous time horizon and at least the first 12 future in-sample horizons are significant at the 1% level. The direct parameters in Table 3 are also essentially equal to the corresponding own parameter from Table 2 which leads to two conclusions. First, in line with results in the empirical spatial literature (e.g.,Autant-Bernard & LeSage, 2011), the direct elasticities contain negligible feedback effects and so in the next subsection the direct returns to scale can be interpreted as own returns. Second, the contemporaneous direct effects essentially persist at the same level over at least the first 12 future in-sample horizons.

In Table 3, for the sample average bank for horizons $0-12$, all but one of the cumulative indirect elasticities for y_1 (net loans and leases), y_2 (securities), y_3 (non-interest income) and p_2 (cost of labor) are significant at the 5% level or lower.[19] Thus, for the sample average bank, a change in p_2 and $y_1 - y_3$ in horizon 0 is associated with symmetric (cumulative) cost spill-ins and spill-outs to and from the bank in future in-sample horizons. It takes three to six future in-sample quarters for these (cumulative) indirect cost impacts to stabilize (or, in other words, for the incremental indirect impacts to die out). To sum up our discussion of Table 3 thus far, if there is in the current period a permanent marginal change for the sample average bank in any of its three outputs or p_2 (but not p_3), there is a notable difference between the persistence of the significant impacts on the bank's direct-own cost and the persistence of the significant impacts on the indirect cost spillovers to and from the bank. That is, it takes notably longer for these impacts on the indirect cost spillovers to die out than it does for the impacts on the direct-own cost. The implication is that, on average, the impacts that relate to a bank's geographical linkages with other banks in the industry exhibit less stability than the impacts that affect individual banks in isolation. This adds to the extant literature as it represents a new type of evidence to support bank regulators' efforts to maintain stability in the banking system.

From Table 3 for horizons $0-12$, we also observe that the significant cumulative indirect impacts for p_2 and $y_1 - y_3$ are exclusively positive/negative and, to different degrees, non-negligible. Consider, for instance, the negative cumulative indirect impacts for y_1 (y_3). These findings indicate that in Table 2, the negative effects of Wy_{1t} and Wy_{1t-1} (Wy_{3t} and Wy_{3t-1}) dominate the positive effects of Wc_t and Wc_{t-1}. The economic interpretation of this involves recognizing that if there is an increase in the $y_1\,(y_3)$ variables of the sample average bank's neighbors in horizon 0, there are two effects in the contemporaneous and future in-sample horizons: (i) the spatial competitive effect of the increase will reduce the sample average bank's cost; and (ii) neighboring banks' costs will increase, which (because neighboring banks' costs are positively spatially dependent on one another) will increase the cost of the sample average bank. Because at the sample mean we observe negative cumulative indirect impacts for y_1 and y_3, for both variables (i) more than offsets (ii).

[19]The exception that is not significant is the indirect elasticity for p_2 in horizon 0.

TABLE 3 Cumulative direct, indirect, and total elasticities for the sample average bank

	y_1 (Net loans & leases)			y_2 (Securities)			y_3 (Non-interest income)			p_2 (Cost of labor)			p_3 (Cost of deposits)		
Horizon, γ	Direct	Indirect	Total	Direct	Indirect	Total	Direct	Indirect	Total	Direct	Indirect	Total	Direct	Indirect	Total
0	0.5605^{c}	−0.1928^{c}	0.3676^{c}	**0.1666**^{c}	0.1344^{c}	0.3011^{c}	**0.1533**^{c}	−0.0745^{c}	0.0788^{c}	**0.6127**^{c}	0.0048	0.6175^{c}	**0.2939**^{c}	−0.0345^{c}	0.2595^{c}
1	**0.5602**^{c}	−0.1684^{c}	0.3918^{c}	0.1669^{c}	0.2076^{c}	0.3745^{c}	0.1532^{c}	−0.0753^{b}	**0.0778**^{c}	0.6128^{c}	0.1056^{b}	0.7184^{c}	0.2939^{c}	0.0013	0.2952^{c}
2	0.5603^{c}	−0.1593^{c}	0.4010^{c}	0.1671^{c}	0.2327^{c}	0.3997^{c}	0.1532^{c}	−0.0754^{b}	0.0777^{b}	0.6130^{c}	0.1406^{c}	0.7536^{c}	0.2940^{c}	0.0138	0.3078^{c}
3	0.5603^{c}	−0.1562^{c}	0.4041^{c}	0.1671^{c}	0.2413^{c}	0.4084^{c}	0.1532^{c}	**-0.0755**^{b}	0.0777^{b}	0.6131^{c}	0.1526^{c}	0.7657^{c}	0.2940^{c}	0.0181	0.3121^{c}
4	0.5603^{c}	**-0.1551**^{c}	0.4052^{c}	0.1671^{c}	0.2443^{c}	0.4114^{c}	0.1532^{c}	−0.0755^{b}	0.0776^{b}	0.6131^{c}	0.1568^{c}	0.7699^{c}	0.2940^{c}	**0.0196**	0.3136^{c}
5	0.5603^{c}	−0.1547^{c}	**0.4056**^{c}	0.1671^{c}	0.2454^{c}	**0.4125**^{c}	0.1532^{c}	−0.0755^{b}	0.0776^{b}	0.6131^{c}	0.1583^{c}	**0.7714**^{c}	0.2940^{c}	0.0202	0.3142^{c}
6	0.5603^{c}	−0.1546^{c}	0.4057^{c}	0.1671^{c}	**0.2457**^{c}	0.4128^{c}	0.1532^{c}	−0.0755^{b}	0.0776^{b}	0.6131^{c}	**0.1588**^{c}	0.7719^{c}	0.2940^{c}	0.0203	0.3144^{c}
7	0.5603^{c}	−0.1545^{c}	0.4057^{c}	0.1671^{c}	0.2459^{c}	0.4130^{c}	0.1532^{c}	−0.0755^{b}	0.0776^{b}	0.6131^{c}	0.1590^{c}	0.7721^{c}	0.2940^{c}	0.0204	0.3144^{c}
8	0.5603^{c}	−0.1545^{c}	0.4058^{c}	0.1671^{c}	0.2459^{c}	0.4130^{c}	0.1532^{c}	−0.0755^{b}	0.0776^{b}	0.6131^{c}	0.1591^{c}	0.7722^{c}	0.2940^{c}	0.0204	**0.3145**^{c}
9	0.5603^{c}	−0.1545^{c}	0.4058^{c}	0.1671^{c}	0.2459^{c}	0.4130^{c}	0.1532^{c}	−0.0755^{b}	0.0776^{b}	0.6131^{c}	0.1591^{c}	0.7722^{c}	0.2940^{c}	0.0204	0.3145^{c}
10	0.5603^{c}	−0.1545^{c}	0.4058^{c}	0.1671^{c}	0.2459^{c}	0.4131^{c}	0.1532^{c}	−0.0755^{b}	0.0776^{b}	0.6131^{c}	0.1591^{c}	0.7722^{c}	0.2940^{c}	0.0204	0.3145^{c}
11	0.5603^{c}	−0.1545^{c}	0.4058^{c}	0.1671^{c}	0.2459^{c}	0.4131^{c}	0.1532^{c}	−0.0755^{b}	0.0776^{b}	0.6131^{c}	0.1591^{c}	0.7722^{c}	0.2940^{c}	0.0205	0.3145^{c}
12	0.5603^{c}	−0.1545^{c}	0.4058^{c}	0.1671^{c}	0.2459^{c}	0.4131^{c}	0.1532^{c}	−0.0755^{b}	0.0776^{b}	0.6131^{c}	0.1591^{c}	0.7722^{c}	0.2940^{c}	0.0205	0.3145^{c}

Notes: a, b and c denote statistical significance at the 10%, 5% and 1% levels, respectively.
The levels of the cumulative elasticities (to 3 dp) that persist over at least the remainder of the first 12 future in-sample quarters are in bold.

We obtain the cumulative total output and input price elasticities for the sample average bank in Table 3 by summing the corresponding direct and indirect elasticities. Therefore, for the sample average bank, the cumulative total elasticities associated with a change in an output or input price of the bank in horizon 0 measure the sum of: (i) the cumulative direct cost impact on the bank and (ii) the symmetric cumulative indirect cost spill-ins and spill-outs to and from the bank. When we compare the total elasticities in Table 3 for the sample average bank, we can see that it takes a different number of quarters for the cumulative total impacts to stabilize, for example, 1 future in-sample quarter for y_3 and 8 for p_3.

5.3 | Spatial returns to scale estimates

In Table 4, for quintiles of the real total asset distribution in 2020:Q4, we present the mean cumulative spatial *RSE* across the study period for time horizons $\gamma = 0,...,4$, that is, the direct, two indirect (spill-in and spill-out), and thus two total *RSE*.[20] In Table 5 we present the corresponding *EPSE* estimates, where in the following discussion we focus on the *EPSE* results because, unlike *RSE*, *EPSE* allow for the possibility that a bank does not lie on a radial ray in the output space. Before we discuss the estimates of the spatial scale economies, we note three things. First, as we report spatial scale economies for subsets of the sample, the asymmetric indirect spill-in and spill-out parameters lead to asymmetric indirect *RSE* (*EPSE*). Second, as for time horizon γ, the two total parameters for a variable are the sum of the direct and indirect spill-in/spill-out parameters, it follows that the two total *RSE* are the sum of the direct and indirect spill-in/spill-out *RSE*. In contrast, for horizon γ, the two total *EPSE* are not the sum of the direct and indirect spill-in/spill-out *EPSE*, as these five spatial *EPSE* measures are ratios with different denominators. Although it is not therefore possible to decompose the two total *EPSE* into direct and indirect *EPSE*, direct and indirect returns to scale are inherent within these total *EPSE* measures. Third, we noted above all the direct parameters for the first-order output and input prices in Table 3 are approximately equal to the corresponding own parameter in Table 2. This indicates that there is very little feedback within the direct parameters and means that the direct returns to scale can be interpreted as own measures.

All the estimates of the direct *EPSE* in Table 5 are less than 1, that is, increasing own returns to scale. There is very little (if any) change in the magnitudes of these direct returns from one horizon to the next, which indicates that the contemporaneous returns persist over future periods. Overlooking for the moment the geographical linkages between banks and considering banks in isolation, this suggests that banks in each quintile are, on average, sub-optimally small in the current period and continue to be in future periods. For each quintile the contemporaneous and dynamic direct *EPSE* in Table 5 are therefore consistent with one another and suggest that, on average, it would be optimal for banks to upsize in the current period and to remain at their new sizes in future periods. This situation is the simpler case where the implications of the contemporaneous and dynamic *EPSE* for the optimal size of a bank are consistent with another. As we will see further in this discussion when we consider the combined effect of banks in isolation and their geographical linkages, we report cases where the contemporaneous and dynamic total *EPSE* for a quintile are at odds with one another, which is when the implications for the optimal bank size become more complex. We first though discuss the results for the indirect spill-in and spill-out returns to scale, which involves overlooking banks in isolation and considering only the impact of their geographical linkages.

To aid the discussion of the indirect returns to scale in Tables 4 and 5, we first focus on their interpretation. If a bank's cumulative indirect spill-out returns are negative (positive) in horizon γ, this indicates that an increase in the bank's outputs in the current period is associated with, on average, a negative (positive) cost spill-out to the other banks in the sample (cumulated over the γ horizons, where γ can be the contemporaneous quarter or a future one). Negative indirect spill-out returns to scale are consistent with an increase in the intensity of spatial competition

[20] A number of banks move between size quintiles over the study period. To make valid comparisons, we must ensure that these banks do not feature in the results for different quintiles. Accordingly, the returns to scale results are for size quintiles in a particular period. This period is the final quarter as this reflects the full evolution of movements between size quintiles over the sample.

TABLE 4 Cumulative spatial ray-scale economies for bank size quintiles

Horizon, γ		Quintile 1	Quintile 2	Quintile 3	Quintile 4	Quintile 5
0	$RSE^{Dir}_{i,\gamma}$	0.871^{*}_{a}	0.873^{*}_{a}	0.877^{*}_{a}	0.882^{*}_{a}	0.899^{*}_{a}
	$RSE^{Ind}_{In,i,\gamma}$	-0.095^{*}_{a}	-0.090^{*}_{a}	-0.089^{*}_{a}	-0.134^{*}_{a}	-0.226^{*}_{a}
	$RSE^{Ind}_{Out,i,\gamma}$	-0.018^{*}_{a}	-0.020^{*}_{a}	-0.034^{*}_{a}	-0.062^{*}_{a}	-0.451^{*}_{a}
	$RSE^{Tot}_{In,i,\gamma}$	0.775^{*}_{a}	0.783^{*}_{a}	0.788^{*}_{a}	0.748^{*}_{a}	0.674^{*}_{a}
	$RSE^{Tot}_{Out,i,\gamma}$	0.852^{*}_{a}	0.853^{*}_{a}	0.843^{*}_{a}	0.819^{*}_{a}	0.448^{*}_{a}
1	$RSE^{Dir}_{i,\gamma}$	0.871^{*}_{a}	0.873^{*}_{a}	0.877^{*}_{a}	0.882^{*}_{a}	0.899^{*}_{a}
	$RSE^{Ind}_{In,i,\gamma}$	-0.028^{*}_{a}	-0.020^{*}_{a}	-0.022^{*}_{a}	-0.035^{*}_{a}	-0.029^{*}_{a}
	$RSE^{Ind}_{Out,i,\gamma}$	-0.007^{*}_{a}	-0.007^{*}_{a}	-0.008^{*}_{a}	-0.013^{*}_{a}	-0.018^{*}_{a}
	$RSE^{Tot}_{In,i,\gamma}$	0.842^{*}_{a}	0.853^{*}_{a}	0.855^{*}_{a}	0.847^{*}_{a}	0.871^{*}_{a}
	$RSE^{Tot}_{Out,i,\gamma}$	0.863^{*}_{a}	0.867^{*}_{a}	0.869^{*}_{a}	0.868^{*}_{a}	0.881^{*}_{a}
2	$RSE^{Dir}_{i,\gamma}$	0.871^{*}_{a}	0.873^{*}_{a}	0.877^{*}_{a}	0.882^{*}_{a}	0.900^{*}_{a}
	$RSE^{Ind}_{In,i,\gamma}$	-0.005^{*}_{a}	0.004^{*}_{a}	0.002^{*}_{a}	0.002_{a}	0.039^{*}_{a}
	$RSE^{Ind}_{Out,i,\gamma}$	-0.002^{*}_{a}	-0.001^{*}_{a}	0.002^{*}_{a}	0.006^{*}_{a}	0.130^{*}_{a}
	$RSE^{Tot}_{In,i,\gamma}$	0.865^{*}_{a}	0.877^{*}_{a}	0.879^{*}_{a}	0.883^{*}_{a}	0.939^{*}_{a}
	$RSE^{Tot}_{Out,i,\gamma}$	0.868^{*}_{a}	0.872^{*}_{a}	0.879^{*}_{a}	0.887^{*}_{a}	1.030^{*}_{b}
3	$RSE^{Dir}_{i,\gamma}$	0.871^{*}_{a}	0.873^{*}_{a}	0.877^{*}_{a}	0.882^{*}_{a}	0.901^{*}_{a}
	$RSE^{Ind}_{In,i,\gamma}$	0.003^{*}_{a}	0.012^{*}_{a}	0.010^{*}_{a}	0.014^{*}_{a}	0.062^{*}_{a}
	$RSE^{Ind}_{Out,i,\gamma}$	-0.001^{*}_{a}	0.001^{*}_{a}	0.006^{*}_{a}	0.012^{*}_{a}	0.180^{*}_{a}
	$RSE^{Tot}_{In,i,\gamma}$	0.873^{*}_{a}	0.885^{*}_{a}	0.887^{*}_{a}	0.896^{*}_{a}	0.963^{*}_{a}
	$RSE^{Tot}_{Out,i,\gamma}$	0.870^{*}_{a}	0.874^{*}_{a}	0.883^{*}_{a}	0.894^{*}_{a}	1.081^{*}_{b}
4	$RSE^{Dir}_{i,\gamma}$	0.871^{*}_{a}	0.873^{*}_{a}	0.877^{*}_{a}	0.882^{*}_{a}	0.901^{*}_{a}
	$RSE^{Ind}_{In,i,\gamma}$	0.005^{*}_{a}	0.015^{*}_{a}	0.013^{*}_{a}	0.018^{*}_{a}	0.071^{*}_{a}
	$RSE^{Ind}_{Out,i,\gamma}$	0.000_{a}	0.002^{*}_{a}	0.007^{*}_{a}	0.014^{*}_{a}	0.198^{*}_{a}
	$RSE^{Tot}_{In,i,\gamma}$	0.876^{*}_{a}	0.888^{*}_{a}	0.890^{*}_{a}	0.900^{*}_{a}	0.971^{*}_{a}
	$RSE^{Tot}_{Out,i,\gamma}$	0.870^{*}_{a}	0.875^{*}_{a}	0.884^{*}_{a}	0.896^{*}_{a}	1.099^{*}_{b}

Notes: * denotes significantly different from zero at the 5% level. *a* denotes significantly less (greater) than 1 (−1) at the 5% level for positive (negative) returns. *b* denotes significantly greater (less) than 1 (−1) at the 5% level for positive (negative) returns.

between banks. We say this because when a bank's outputs increase in the current period, the average negative cost spill-out across the other banks in horizon γ is consistent with, on average, downward spatial competitive pressure on the other banks' costs. Positive indirect spill-out returns to scale indicate that there is positive spatial dependence among the banks, which is consistent with banks being impacted by common economic phenomena, that is, headline changes in economies (national, regional, etc.) and common monetary and regulatory policies.

Interpreting a bank's cumulative indirect spill-in returns to scale in horizon γ requires careful consideration. This is because if these returns are negative (positive), then an increase in the bank's outputs in the current period is associated with a negative (positive) cumulative cost spill-in to the bank (from the other banks) in horizon γ. Negative

TABLE 5 Cumulative spatial expansion-path scale economies for bank size quintiles

Horizon, γ		Quintile 1	Quintile 2	Quintile 3	Quintile 4	Quintile 5
0	$EPSE^{Dir}_{i,\gamma}$	0.946^{*}	0.988^{*}	0.880^{*}_{a}	0.907^{*}	0.912^{*}
	$EPSE^{Ind}_{In,i,\gamma}$	0.899^{*}	0.990^{*}	0.983^{*}	0.950^{*}	1.018^{*}_{b}
	$EPSE^{Ind}_{Out,i,\gamma}$	0.969^{*}	0.953^{*}	0.997^{*}	1.093^{*}	1.019^{*}_{b}
	$EPSE^{Tot}_{In,i,\gamma}$	0.957^{*}	0.975^{*}	0.945^{*}_{a}	0.967^{*}	0.957^{*}_{a}
	$EPSE^{Tot}_{Out,i,\gamma}$	0.998^{*}	0.993^{*}	0.983^{*}	0.975^{*}	0.988^{*}
1	$EPSE^{Dir}_{i,\gamma}$	0.948^{*}	0.988^{*}	0.879^{*}_{a}	0.923^{*}	0.911^{*}
	$EPSE^{Ind}_{In,i,\gamma}$	0.976^{*}	0.952^{*}	0.964^{*}	0.892^{*}	0.959^{*}
	$EPSE^{Ind}_{Out,i,\gamma}$	0.975^{*}	0.968^{*}	0.823^{*}_{a}	0.991^{*}	1.152^{*}_{b}
	$EPSE^{Tot}_{In,i,\gamma}$	0.960^{*}	0.943^{*}_{a}	0.861^{*}	0.990^{*}	0.935^{*}_{a}
	$EPSE^{Tot}_{Out,i,\gamma}$	0.995^{*}	0.977^{*}	0.969^{*}	0.953^{*}_{a}	0.985^{*}
2	$EPSE^{Dir}_{i,\gamma}$	0.949^{*}	0.988^{*}	0.879^{*}_{a}	0.927^{*}	0.911^{*}
	$EPSE^{Ind}_{In,i,\gamma}$	0.962^{*}_{a}	0.992^{*}	0.962^{*}	0.927^{*}	0.856^{*}
	$EPSE^{Ind}_{Out,i,\gamma}$	0.984^{*}	0.969^{*}_{a}	0.932^{*}_{a}	0.939^{*}_{a}	1.003^{*}
	$EPSE^{Tot}_{In,i,\gamma}$	0.988^{*}	0.961^{*}	0.895^{*}	0.933^{*}_{a}	0.995^{*}
	$EPSE^{Tot}_{Out,i,\gamma}$	0.983^{*}	0.867^{*}_{a}	0.941^{*}	0.887^{*}	1.005^{*}
3	$EPSE^{Dir}_{i,\gamma}$	0.949^{*}	0.988^{*}	0.879^{*}_{a}	0.928^{*}	0.911^{*}
	$EPSE^{Ind}_{In,i,\gamma}$	0.940^{*}	0.894^{*}_{a}	0.999^{*}	0.947^{*}	0.822^{*}_{a}
	$EPSE^{Ind}_{Out,i,\gamma}$	0.937^{*}_{a}	0.975^{*}	0.948^{*}_{a}	0.924^{*}	1.001^{*}
	$EPSE^{Tot}_{In,i,\gamma}$	0.963^{*}	0.908^{*}_{a}	0.944^{*}	0.935^{*}_{a}	0.990^{*}
	$EPSE^{Tot}_{Out,i,\gamma}$	0.942^{*}	0.963^{*}	0.900^{*}_{a}	0.878^{*}	1.008^{*}
4	$EPSE^{Dir}_{i,\gamma}$	0.949^{*}	0.988^{*}	0.879^{*}_{a}	0.929^{*}	0.911^{*}
	$EPSE^{Ind}_{In,i,\gamma}$	0.951^{*}	0.881^{*}_{a}	0.905^{*}	0.932^{*}	0.893^{*}_{a}
	$EPSE^{Ind}_{Out,i,\gamma}$	0.954^{*}_{a}	0.970^{*}	0.988^{*}	0.889^{*}	1.008^{*}
	$EPSE^{Tot}_{In,i,\gamma}$	0.948^{*}	0.924^{*}_{a}	0.978^{*}	0.846^{*}	0.981^{*}
	$EPSE^{Tot}_{Out,i,\gamma}$	0.945^{*}	0.922^{*}	0.883^{*}_{a}	0.961^{*}	1.010^{*}

Notes: At the 5% level, * denotes significantly different from zero, *a* denotes significantly less than 1, and *b* denotes significantly greater than 1.

indirect spill-in returns to scale are not counterintuitive, as an increase in the outputs of all but one bank in the sample in the current period is consistent with spatial competition intensifying for the remaining bank, leading to downward pressure on its costs. To reflect this pressure, attached to an increase in the outputs of the remaining bank in the current period is a negative cost spill-in to the bank in horizon γ. We now turn to positive indirect spill-in returns and the role of positive spatial dependence. In this case, if there is an increase in the outputs of all but one bank in the current period, attached to an increase in the outputs of the remaining bank in the current period is a positive cost spill-in to the bank in horizon γ.

We noted above that we focus on the *EPSE* results because, unlike *RSE*, *EPSE* allows for the possibility that a bank does not lie on a radial ray in the output space. Although we observe differences between corresponding direct

RSE and *EPSE* in Tables 4 and 5, of the three types of *RSE* and *EPSE* results in these tables (i.e., direct, indirect, and total), we have the strongest preference for indirect *EPSE* over indirect *RSE*, as this is where there is the greatest difference between our *EPSE* and *RSE* findings. We conclude therefore that extending own *RSE* and *EPSE* to the spatial setting strengthens the case for the *EPSE* method over the *RSE* approach.

To illustrate, there are several notable differences between the reported cumulative indirect *RSE* and *EPSE*. Specifically, all the cumulative indirect *EPSE* in Table 5 are positive, non-negligible, significantly different from zero, and persistent over at least horizons $0-4$. The cumulative indirect *RSE* in Table 4, however, are a mix of positive and negative estimates. As we discussed above, negative indirect returns are consistent with cost reductions that spill-in/spill-out because of more intense spatial competition, while positive indirect returns are consistent with cost increases that spill-in/spill-out because of the positive spatial dependence among the banks. Moreover, in contrast to the cumulative indirect *EPSE*, for all five quintiles the cumulative indirect *RSE* does not tend to persist at (broadly) similar levels across horizons $0-4$. For instance, the cumulative indirect (spill-in and spill-out) *RSE* for quintile 5 increases sharply from their large negative values in horizon 0 to non-negligible positive values in horizon 4.

Given our preference for the *EPSE* estimates, to ascertain the impact of indirect returns to scale spillovers on where a bank's returns sit relative to the optimal level, we should compare the direct *EPSE* (i.e., when we overlook the indirect returns) with the two total *EPSE*. More specifically, from a bank's point of view it would focus on $EPSE^{Tot}_{In,i,\gamma}$ and not on $EPSE^{Tot}_{Out,i,\gamma}$, as the former incorporates the indirect returns that spill-in to the bank, while the latter accounts for its indirect returns that spill-out to the other banks. $EPSE^{Tot}_{In,i,\gamma}$ would also be of interest to bank regulators who may also find $EPSE^{Tot}_{Out,i,\gamma}$ informative, as the latter measures whether a bank's returns to scale are optimal from the perspective of the cost implications for all the banks in the analysis. Alternatively, by focusing on a bank's indirect *EPSE*, one can analyze the returns that spill in (spill out) to (from) a bank from (to) the other banks and which involve excluding self-impact. This is the approach we use in the next subsection to analyze the spatial interactions of the largest banks.

At this juncture we focus on a bank's point of view and for each quintile compare the mean direct *EPSE* and $EPSE^{Tot}_{In,i,\gamma}$ for horizons $0-4$. This comparison for quintile 1 banks indicates that these direct and total measures have (broadly) similar magnitudes that point to increasing returns. We can therefore conclude that returns to scale spill-ins have little impact on the level of the optimal returns of the mean bank in quintile 1. This can be attributed to there being an insufficient overlap between the branch networks of the banks in this quintile, as these networks are comparatively small in our sample. In other words, on average, there is insufficient spatial interaction between a bank in quintile 1 and the others in the sample. The optimal size implications of the $EPSE^{Tot}_{In,i,\gamma}$ for quintile 1 banks are therefore the same as those discussed above for the direct *EPSE*. That is, for quintile 1 banks the contemporaneous and dynamic total $EPSE^{Tot}_{In,i,\gamma}$ are consistent with one another and suggest that, on average, these banks should upsize in the current period and remain at their new sizes in future periods, which would likely intensify present and future competition in the industry as these banks are the smallest in our sample.[21] This situation is another example of the simpler case where the optimal size implications of the contemporaneous and dynamic *EPSE* are consistent with one another.

To use our results to draw conclusions on how indirect spill-in returns to scale impact where the returns of the mean banks in quintiles $2-4$ sit relative to their optimal levels, we make two remarks about the reported direct *EPSE* and $EPSE^{Tot}_{In,i,\gamma}$. First, although to different degrees the magnitudes of the mean contemporaneous direct and total returns differ, both returns are always less than 1 and their significance is the same, for example, not significantly less than constant returns for quintile 4. In non-spatial and static spatial settings, these types of contemporaneous results would be used to make inferences about the optimal size of a bank. Based only on these results, on average, the banks in quintiles $2-4$ should upsize, although our next remark indicates that in the dynamic spatial setting these inferences are much more complex.

[21]We thank an anonymous reviewer for suggesting us to relate our returns to scale results to the competitive environment.

Second and importantly, in line with the reasonable view that it can take time for spillovers to occur, we can see for quintiles 2 – 4 that it takes some time (one or two quarters) for there to be a significant departure from the contemporaneous total returns, for example, for quintile 4 it takes two quarters before we observe a change in significant increasing returns. These total returns for quintiles 2 – 4 are therefore examples of the complex case where the implications for the optimal size of a bank of the contemporaneous and dynamic total *EPSE* are at odds with one another. This is because such changes in these total returns in subsequent quarters means that a bank can be in situation whereby if it acts on its contemporaneous total *EPSE* by, for example, not changing/increasing its outputs by a particular percentage, this can lead to its returns being suboptimal in the following quarters. To manage these dynamic returns to scale effects, we suggest that a bank should seek to optimize its returns over the time frame of its future plans, which would involve its returns being suboptimal in particular periods within this time frame.

In the above case of quintile 1 banks, the contemporaneous and dynamic $EPSE^{Tot}_{ln,i,\gamma}$ are consistent with one another, so it is much clearer that if, on average, these banks were to upsize in the current period and remain at their new sizes in future periods, this would likely intensify present and future competition in the industry as these banks are the smallest in our sample. However, the contemporaneous and dynamic $EPSE^{Tot}_{ln,i,\gamma}$ for quintiles 2 – 4 are at odds with one another, so it is less clear what the implications would be for the competitive environment if, on average, these banks optimize their returns over the time frame of their future plans. We can conclude though for quintiles 2 – 4 that the types of inferences about the optimal sizes of banks in the non-spatial and static spatial settings, versus those from our dynamic spatial setting, would likely be associated with notable differences in the competitive environments.

For quintile 5, the reported mean direct *EPSE* and $EPSE^{Tot}_{ln,i,\gamma}$ are all less than 1 and these total returns (i.e., when we account for indirect spill-in returns) are (noticeably) larger than the direct *EPSE*. Relatedly, these direct and total *EPSE* are either not significantly less than constant returns or significantly less than 1. The starting point for the discussion of these results is the role of the excessive risk taking by very large U.S. banks in the 2008 crisis. This risk taking was promoted by the TBTF status of these banks, and accordingly this status was a focus of the 2010 Dodd-Frank regulatory reforms. These reforms involved taking steps to avoid a repeat of this risk taking and included tightening bank regulation through, for instance, more stringent liquidity constraints, and establishing a formal process to resolve large bank failures with the intention of no bank being TBTF. However, Fisher and Rosenblum (2012), for example, argue that Dodd-Frank would not prevent TBTF banks and that these reforms should have taken further preventative action by imposing size caps on the largest banks. A key practical insight from the mean direct *EPSE* and $EPSE^{Tot}_{ln,i,\gamma}$ results for quintile 5 is that, on average, size caps would have a negative effect on how efficiently the largest banks use society's resources to provide their products.[22] That said, when assessing whether size caps should be imposed, this cost should be balanced against the benefits associated with preventing banks from becoming TBTF that are outside the scope of our analysis, such as the reduced risk exposure of the largest banks.

In Figure 1, for quintiles 1, 3, and 5 of the real total asset distribution in 2020:Q4, we present the contemporaneous direct *RSE* over the study period. Of the full set of spatial *RSE* results (the direct, indirect spill-in and spill-out, and two total measures for various horizons) we only present the direct *RSE* results for horizon 0 for illustrative purposes. This is because, as we noted above, we have a preference for *EPSE* over *RSE*, and we also observed above from Tables 4 and 5 that there is more similarity between the direct *RSE* and *EPSE* than there is between the corresponding indirect and total measures. Accordingly, in Figure 2 for the same three quintiles and horizons 0 and 1, we present the full set of spatial *EPSE* results over the study period.[23] Note that in both figures the gray vertical bars represent the timing of U.S. recessions.[24] For the three quintiles, Figure 1 indicates that the COVID pandemic had very little impact on the contemporaneous direct (i.e., own) *RSE*. We can see from Figure 2 that this is also the case for the contemporaneous direct *EPSE* for quintile 5. There is some evidence that COVID had more of an impact

[22]This is in line with the conclusion for the largest U.S. banks from Wheelock and Wilson's (2018) analysis of own scale economies (1986:Q4 – 2015:Q4).
[23]All the *EPSE* results in Figure 2 are after outliers have been removed as they accentuate the volatility; that is, the top and bottom 2.5% of bank-year estimates are dropped.
[24]The dates of the recessions are from the FRED database.

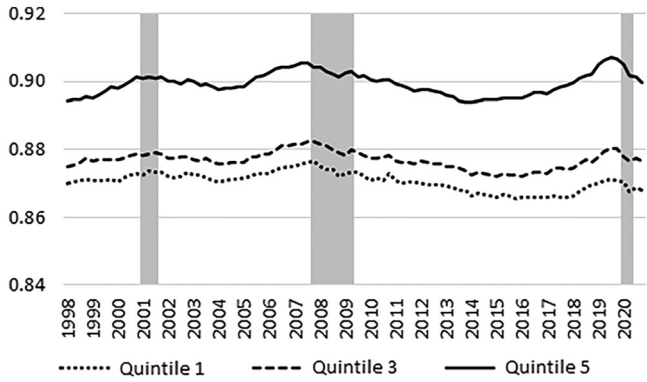

FIGURE 1 Quarterly contemporaneous direct ray-scale economies

on the contemporaneous direct *EPSE* for quintile 1, although the more noticeable effects of COVID are on the indirect *EPSE* which is what we discuss in detail next. Our analysis, however, only considers the impact of the 2020 portion of the pandemic. An area for future research therefore is to analyze the impact of the entirety of the pandemic on banks' scale economies.

Interestingly, for each of the three quintiles in Figure 2, we can see that there is hardly any asymmetry between the contemporaneous indirect spill-in and spill-out *EPSE*. There is, however, evidence of asymmetry between the two indirect *EPSE* for horizon 1, which is much more marked for quintile 5. There is also evidence of differences between the results for larger and smaller banks during the 2008 crisis and the COVID pandemic. During the COVID pandemic, we can see that the two contemporaneous indirect *EPSE* for quintile 1 declined further below 1, whereas for quintile 5 they increased further above 1. In terms of the cost spill-in and spill-out interactions of a quintile 1 (quintile 5) bank with the other banks in the sample, this suggests that, on average, the pandemic led to quintile 1 (quintile 5) banks becoming suboptimally smaller (larger). This is consistent with large banks providing most of the required funding to firms in the U.S. when they turned to the banks for the liquidity provision using pre-existing lines of credit during the pandemic (Li et al., 2020). Relatedly, we can see from Figure 2 that during the 2008 crisis there are differences between the direct (indirect) *EPSE* results for quintiles 1 and 5, while differences are also observed between these results and the corresponding result for the same quintile during the pandemic. Differences between these *EPSE* results are not unexpected and point to the inadequacy of general purpose responses in such situations because although the crisis and the pandemic were both detrimental shocks these shocks were very different in nature, while the business models of the smallest and largest banks are also very different.[25]

Now we turn to consider the findings following the 2008 crisis. For quintiles 1 and 3 and horizons 0 and 1, there are sustained declines in the two indirect *EPSE* measures. Conversely, for quintile 5 there are steady / irregular increases in these indirect measures. We can also see for quintile 5 and horizons 0 and 1 that post-crisis there are upward trends in the direct and two total *EPSE*. Such increases coincide with a period where there were notable increases in the sizes of banks that are among the largest in the industry. Despite these size increases, the own scale economies that Wheelock and Wilson (2018) report suggest that the biggest U.S. banks do not tend to be suboptimally large. Their study period ends in 2015:Q4, and at this point in our sample we can see for quintile 5 that the contemporaneous direct *EPSE* and one of the contemporaneous total *EPSE* are slightly above 1 and increase gradually in the following few years. For this period in isolation, this suggests that, on average, appropriate small size caps on the largest banks would improve how efficiently they use society's resources to provide their products. This conclusion does not also apply to the next few years of our sample because the contemporaneous direct and total

[25]We thank an anonymous reviewer for highlighting this finding from Figure 2.

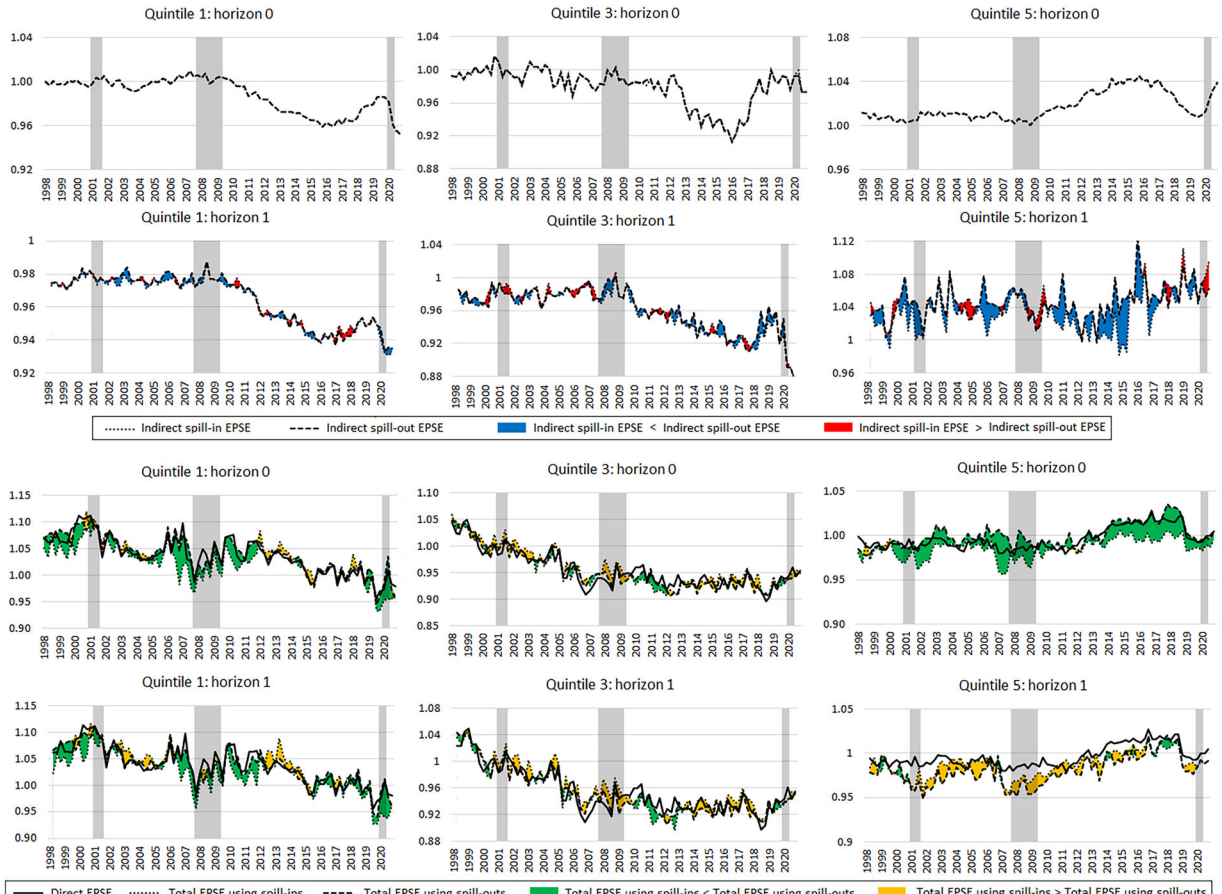

FIGURE 2 Quarterly cumulative spatial expansion-path scale economies

EPSE for quintile 5 fall below 1. Given we find that there can be short-term changes in the implications of size caps for banks, if regulators use scale economies to inform such policy decisions, we suggest they consider these returns over a sufficiently long period and have a similarly long-term policy view.

5.4 | Spatial interactions of the largest banks

We focus on the banks in quintile 5 of the real total assets distribution in 2020:Q4 as this includes the TBTF banks. From this quintile, Table 6 presents the top five ranked banks with the largest positive estimates of eight *EPSE*measures of the spatial interactivity of a bank's scale economies. The eight measures are means over the sample and are a mix of contemporaneous and dynamic measures. (i) and (ii) are the contemporaneous indirect spill-out and spill-in *EPSE*, respectively. (iii) and (iv) are the net contemporaneous indirect spill-out and spill-in *EPSE*, which are (i) minus and (ii) vice versa. (v) is the dynamic indirect spill-out *EPSE*, which is the mean of this measure over the next four in-sample quarters. (vi) is the dynamic indirect spill-in *EPSE*, which is calculated in the same way as (v). (vii) and (viii) are the net dynamic indirect spill-out and spill-in *EPSE*, which are (v) minus and (vi) vice versa.

To set the scene for the discussion of the results in Table 6 we note two things. First, we can interpret the banks with a high-ranking contemporaneous and dynamic indirect spill-out *EPSE* as those who when their outputs increase in the current period *initiate* the spill-out of the largest cost returns to the other banks in the sample in current and future periods. Conversely, we can interpret the banks with a high-ranking contemporaneous and dynamic indirect spill-in *EPSE* as those who when their outputs increase in the current period are the *recipients* of the largest cost returns that spill-in from other banks in current and future periods. Second, we report the banks with the highest-ranking net contemporaneous and dynamic measures to examine which banks have the highest-ranking asymmetries between their indirect spill-out and spill-in *EPSE*.

The general finding is that the top 5 ranked banks for the eight measures in Table 6 include a wide range of banks, a small number of which are global systemically important banks (G-SIBs) (Financial Stability Board, 2019). We elaborate on this in the discussion of the following three specific findings from Table 6. First, a number of the banks with a high-ranking (net) contemporaneous indirect spill-in and/or spill-out *EPSE* have geographically concentrated branches and/or specialize in particular activities. For example, all of the branches of the Independent Bank are in Michigan, while the Silicon Valley Bank focuses on funding hi-tech businesses and although it has a very small number of branches it has the largest local deposits in Silicon Valley. Such high-ranking indirect *EPSE* may be because the geographical and operational focus of the banks is associated with higher-quality service spillovers that lead to relatively high-cost spillovers.

Second, JPMorgan Chase and Wells Fargo have high-ranking net dynamic indirect spill-out *EPSE* measures, whereas there are other G-SIBs with much lower corresponding rankings, namely, Bank of America and Citibank that are ranked 20 and 22, respectively. This indicates that only certain G-SIBs are among the initiators of the largest net cost returns that spill-out to the other banks in the sample in future periods. Third, of the banks included in the *Comprehensive Capital Analysis and Review* (CCAR) (Federal Reserve Board, 2019), which is a group of banks where a constituent bank has the capability to influence the domestic U.S. banking industry, only Capital One, Fifth Third, and the Discover Bank have high-ranking (net) dynamic spill-in *EPSE*. This suggests that only certain CCAR banks are among the recipients of the largest cost returns that spill-in from the other banks in future periods.

In Figure 3, for the quintile 5 banks, we map their net dynamic indirect spill-out *EPSE* onto their 2020 branch networks. Note that the corresponding map of the net dynamic indirect spill-in *EPSE* is the reverse of Figure 3; that is, the bank branch networks in red in Figure 3 are the banks with the lowest net dynamic indirect spill-in *EPSE*. We can see from Figure 3 that banks with the highest net dynamic indirect spill-out *EPSE* have branches that cluster in areas such as New York City and the surrounding area and along the West Coast. We now turn to a comparison of Figure 3 and the 2020 branch networks of JPMorgan Chase and Wells Fargo in Figure 4. Consistent with these banks having net dynamic indirect spill-out *EPSE* that are in the top three in our sample, this comparison indicates

TABLE 6 Highest-ranked quintile 5 banks for spill-out and spill-in EPSE

Bank	Real total assets (millions)	Number of branches	Bank	Real total assets (millions)	Number of branches
Contemporaneous spill-out EPSE			**Net contemporaneous spill-out EPSE**		
1. Silicon Valley Bank	99,609 (20)	5	1. Silicon Valley Bank	99,609 (20)	5
2. Renasant Bank	13,052 (77)	172	2. Bank Ozk	23,767 (51)	240
3. Bank Ozk	23,767 (51)	240	3. Centennial Bank	14,321 (72)	170
4. Centennial Bank	14,321 (72)	170	4. Simmons Bank	19,520 (60)	233
5. Sterling National	26,022 (47)	80	5. Eastern Bank	13,958 (74)	89
Contemporaneous spill-in EPSE			**Net contemporaneous spill-in EPSE**		
1. Silicon Valley Bank	99,609 (20)	5	1. Citizens Business	12,612 (79)	59
2. Renasant Bank	13,052 (77)	172	2. Fifth Third Bank	177,776 (12)	1,137
3. Bank Ozk	23,767 (51)	240	3. Independent Bank	15,529 (67)	96
4. Centennial Bank	14,321 (72)	170	4. UMB Bank	28,854 (45)	95
5. Sterling National	26,022 (47)	80	5. New York Mellon	367,435 (9)	36
Dynamic spill-out EPSE			**Net dynamic spill-out EPSE**		
1. Centennial Bank	14,321 (72)	170	1. People's United	55,314 (30)	421
2. Cadence Bank	16,363 (64)	102	2. JPMorgan Chase	2,647,111 (1)	4,979
3. Arvest Bank	21,311 (57)	273	3. Wells Fargo Bank	1,546,824 (3)	5,410
4. Synovus Bank	47,535 (32)	292	4. Centennial Bank	14,321 (72)	170
5. First Horizon Bank	73,392 (27)	269	5. Cadence Bank	16,363 (64)	102
Dynamic spill-in EPSE			**Net dynamic spill-in EPSE**		
1. Capital One Bank	318,080 (10)	451	1. Capital One Bank	318,080 (10)	451
2. Fifth Third Bank	177,776 (12)	1,137	2. Fifth Third Bank	177,776 (12)	1,137
3. First Horizon Bank	73,392 (27)	269	3. Wilmington Savings	12,511 (80)	95
4. Wilmington Savings	12,511 (80)	95	4. Discover Bank	97,418 (21)	2
5. Discover Bank	97,418 (21)	2	5. City Nat. Bank of FL	16,289 (65)	32

Notes: Real total assets are for 2020:Q4 with the rankings in parentheses. The number of branches is the mid-year measure in 2020.

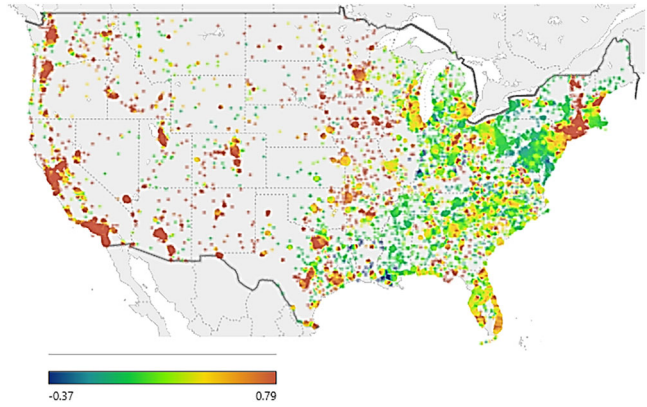

FIGURE 3 Branch locations and net dynamic spill-out EPSE for quintile 5 banks

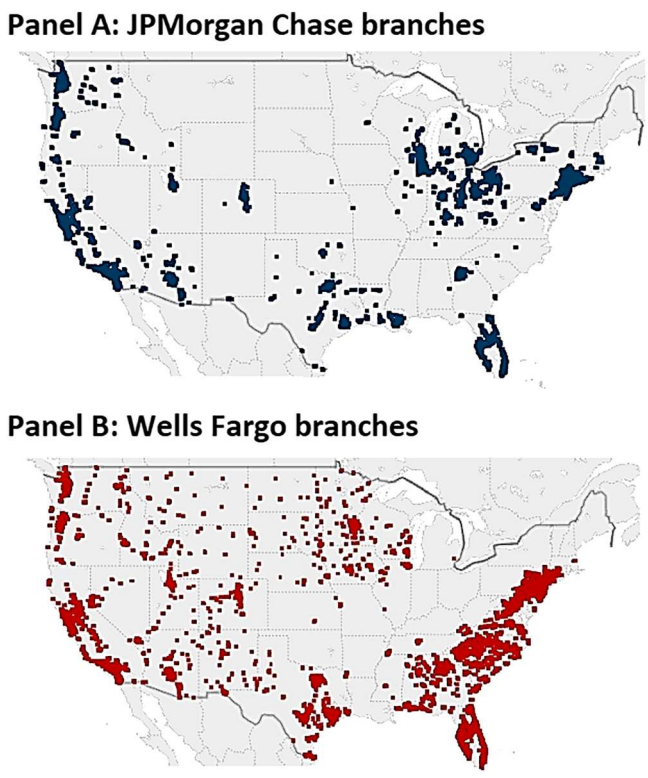

FIGURE 4 JPMorgan Chase and Wells Fargo mid-year 2020 branches

that they have clusters of branches in areas along the West Coast and in New York City and the surrounding area. Interestingly, in Figure 4 there are parts of the U.S. where Wells Fargo (JPMorgan Chase) has a notable branch presence, but where Figure 3 suggests that banks operating in these areas tend to have relatively low net dynamic indirect spill-out *EPSE*, for example, Wells Fargo's branches in the Southeast. This is because in such areas the relatively

low net dynamic indirect spill-out *EPSE* of other banks more than offsets the high value of this measure for Wells Fargo (JPMorgan Chase).

6 | CONCLUDING SUMMARY WITH INSIGHTS FOR BANKS AND REGULATORS

Spatial data science methods are a toolkit that is specifically designed to, among other things, account for the impacts of geographical links, namely, who one's neighbors are and the nature of the links between neighbors. By drawing on this spatial dependence in the data, this toolkit uses interneighbor spillovers to measure the impacts of these links. This toolkit has been widely applied to regions, states, and cities because of their geographical connectedness, for example, a common border or their relative close proximity. These methods are also well suited to banks to analyze the impacts of the overlapping geography of their branch networks. To this end, we provide an accessible presentation of the approach to apply spatial data science methods to analyze these impacts. In particular, our approach uses a dynamic spatial model because it is reasonable to think that it can take some time for geographical spillovers to occur. We present this approach in terms of our empirical application to calculate indirect spill-in and spill-out returns to scale using a dynamic spatial cost function. Areas for further work therefore include empirical applications of our approach to returns of scale of firms in other industries; returns to scale of U.S. banks for the entirety of the pandemic and beyond and from other functions, such as dynamic spatial revenue and alternative profit functions; and returns to scale of banks in other countries. As our approach is general, it could also be applied to analyze the persistence of spillovers in models of other bank variables where the geography of activities is regarded as an important determinant, including market value, loan quality measures, loan growth, deposits per branch, small business loans per branch, and bank profitability. Moreover, in terms of applied methodology, our model could be extended to allow for the possibility that there is some inefficiency in the operations of the units (banks, regions, etc.), as we are not aware of a study that presents a dynamic spatial stochastic frontier model.

We draw on two appealing features of the spatial model we propose. First, the contemporaneous and dynamic indirect spill-in and spill-out elasticities measure how the impacts of the geography of activities are manifested within the impacts of bank variables, such as in our application within the impacts of bank outputs. Second, these indirect elasticities can be used to calculate informative post-estimation measures, such as the contemporaneous and dynamic returns to scale spill-ins and spill-outs we compute. We then use these measures to ascertain which banks are the top-ranked initiators of the biggest contemporaneous and dynamic returns to scale spill-outs, and which are the top-ranked recipients of the biggest contemporaneous and dynamic spill-ins.

We compute cost-oriented contemporaneous and dynamic spatial returns to scale using quarterly panel data for medium-sized and large U.S. banks. Three main findings are as follows. First, we observe that the most noticeable impacts of the 2020 portion of the COVID pandemic are on banks' indirect spill-in and spill-out *EPSE*, while there is relatively little impact on their direct-own *EPSE*. We observe a clear difference between the impacts of the pandemic on the two contemporaneous indirect *EPSE* for smaller and larger banks. For quintile 1 of the bank size distribution, the pandemic led to these two *EPSE* measures declining further below 1, while for quintile 5 they increased further above 1. From the perspective of the cost spill-in and spill-out interactions of a quintile 1 (quintile 5) bank with the other banks in the sample, this suggests that, on average, the pandemic led to quintile 1 (quintile 5) banks becoming suboptimally smaller (larger). This is consistent with quintile 5 banks having more resources to absorb the impact of the COVID-induced changes in the banking environment. Second, we find that a number of banks with a high-ranking contemporaneous indirect spill-in and/or spill-out *EPSE* have geographically concentrated branches and/or specialize in particular activities, for example, Silicon Valley Bank and Independent Bank. Such high-ranking indirect *EPSE* may be because the geographical and operational focus of the banks is associated with higher-quality service spillovers and thus relatively high-cost spillovers. Third, we observe that certain G-SIBs (JPMorgan Chase and Wells

Fargo) have a high-ranking net dynamic indirect spill-out *EPSE*, whereas there are other G-SIBs (Bank of America and Citibank) with much lower rankings for this measure.

These second and third findings highlight to regulators the leading roles that certain distinctive large banks and certain G-SIBs play in bank cost interdependence. These findings also highlight how the nature of these roles can differ, namely, a contemporaneous role or a dynamic one, an absolute role or a net one, and finally, being one of the initiators of the largest cost spill-outs, or one of the recipients of the largest cost spill-ins. From a policy perspective, ongoing bank stress testing by regulators will include the G-SIBs and CCAR banks, but regulators may consider extending this testing to include the other large banks that we find are prominent in bank cost interdependence.

Finally, we note two limitations of our study. First, for the scope of our analysis to be manageable, we use only a cost function to analyze where banks' returns to scale sit relative to the optimal levels. Although this is the most common approach to analyze bank returns to scale, returns to scale can also be analyzed using other functions, such as revenue and alternative profit functions. This then raises the issue of whether the findings on returns to scale are robust across different functions. Second and relatedly, bank returns to scale from a cost function reflect how efficiently a bank uses society's resources to provide its products, but this resource efficiency will not be the only objective of stakeholders. Changes in revenue and profit-oriented returns to scale would be of interest to bank shareholders, and policymakers would also be interested in whether such changes have implications for industry consolidation.

ORCID

Anthony J. Glass https://orcid.org/0000-0002-5984-7666
Karligash Kenjegalieva https://orcid.org/0000-0001-7323-077X

REFERENCES

Aguirregabiria, V., Clark, R., & Wang, H. (2016). Diversification of geographic risk in retail bank networks: Evidence from bank expansion after the Riegle-Neal Act. *RAND Journal of Economics*, *47*, 529–572.

Algeri, C., Anselin, L., Forgione, A. F., & Migliardo, C. (2022). Spatial dependence in the technical efficiency of local banks. *Papers in Regional Science*, *101*, 685–716.

Anselin, L. (1988). *Spatial Econometrics: Methods and Models*. Dordrecht: Kluwer.

Autant-Bernard, C., & LeSage, J. P. (2011). Quantifying knowledge spillovers using spatial econometric models. *Journal of Regional Science*, *51*, 471–496.

Berger, A. N., & DeYoung, R. (2001). The effects of geographic expansion on bank efficiency. *Journal of Financial Services Research*, *19*, 163–184.

Berger, A. N., & DeYoung, R. (2006). Technological progress and the geographic expansion of the banking industry. *Journal of Money, Credit and Banking*, *38*, 1483–1513.

Berger, A. N., Hanweck, G. A., & Humphrey, D. B. (1987). Competitive viability in banking- scale, scope and product mix economies. *Journal of Monetary Economics*, *20*, 501–520.

Berger, A. N., & Mester, L. J. (1997). Inside the black box: What explains differences in the efficiencies of financial institutions? *Journal of Banking and Finance*, *21*, 895–947.

Berger, A. N., & Roman, R. A. (2017). Did saving Wall Street really save Main Street? The real effects of TARP on local economic conditions. *Journal of Financial and Quantitative Analysis*, *52*, 1827–1867.

Boarnet, M. G., & Glazer, A. (2002). Federal grants and yardstick competition. *Journal of Urban Economics*, *52*, 53–64.

Chu, Y., Deng, S., & Xia, C. (2020). Bank geographic diversification and systemic risk. *Review of Financial Studies*, *33*, 4811–4838.

Clark, J. A., & Siems, T. F. (2002). X-efficiency in banking: Looking beyond the balance sheet. *Journal of Money, Credit and Banking*, *34*, 987–1013.

Corrado, L., & Fingleton, B. (2012). Where is the economics in spatial econometrics? *Journal of Regional Science*, *52*, 210–239.

Debarsy, N., Ertur, C., & LeSage, J. P. (2012). Interpreting dynamic space-time panel data models. *Statistical Methodology*, *9*, 158–171.

Degryse, H., & Ongena, S. (2005). Distance, lending relationships and competition. *Journal of Finance*, *60*, 231–266.

Deng, S., & Elyasiani, E. (2008). Geographic diversification, bank holding company value, and risk. *Journal of Money, Credit and Banking*, *40*, 1217–1238.

DeYoung, R., Klier, T., & McMillen, D. (2004). The changing geography of the U.S. banking industry. *The Industrial Geographer*, *2*, 29–48.

Dick, A. A. (2006). Nationwide branching and its impact on market structure, quality, and bank performance. *Journal of Business*, *79*, 567–592.

Druska, V., & Horrace, W. C. (2004). Generalized moments estimation for spatial panel data: Indonesian rice farming. *American Journal of Agricultural Economics*, *86*, 185–198.

Elhorst, J. P. (2009). Spatial panel data models. In Fischer, M. M., & Getis, A. (Eds.), *The Handbook of Applied Spatial Analysis* (pp. 377–408). Springer.

Federal Reserve Board (2019). Comprehensive Capital Analysis and Review Quantitative Results, 2013 – 2019. Available at: https://www.federalreserve.gov/supervisionreg/files/public_results_CCAR.csv (last accessed on 03/31/22).

Financial Stability Board (2019). 2019 list of global systemically important banks (G-SIBs). Available at: https://www.fsb.org/wp-content/uploads/P221119-1.pdf (last accessed on 03/31/22).

Fisher, R. W., & Rosenblum, H. (2012). Vanquishing too Big to Fail. In *2012 Annual Report of the Federal Reserve Bank of Dallas* (pp. 5–10). Federal Reserve Bank, Dallas, TX.

Garrett, T. A., & Marsh, T. L. (2002). The revenue impacts of cross-border lottery shopping in the presence of spatial autocorrelation. *Regional Science and Urban Economics*, *32*, 501–519.

Glass, A. J., Kenjegaliev, A., & Kenjegalieva, K. (2020). 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., & Kenjegalieva, K. (2019). A spatial productivity index in the presence of efficiency spillovers: Evidence for U.S. banks, 1992–2015. *European Journal of Operational Research*, *273*, 1165–1179.

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

Glass, A. J., Kenjegalieva, K., & Paez-Farrell, J. (2013). Productivity growth decomposition using a spatial autoregressive frontier model. *Economics Letters*, *119*, 291–295.

Glass, A. J., Kenjegalieva, K., & Sickles, R. C. (2014). Estimating efficiency spillovers with state level evidence for manufacturing in the US. *Economics Letters*, *123*, 154–159.

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

Glass, A. J., Kenjegalieva, K., & Weyman-Jones, T. (2020). The effect of monetary policy on bank competition using the Boone index. *European Journal of Operational Research*, *282*, 1070–1087.

Goetz, M. R., Laeven, L., & Levine, R. (2013). Identifying the valuation effects and agency costs of corporate diversification: Evidence from the geographic diversification of US banks. *Review of Financial Studies*, *26*, 1787–1823.

Goetz, M. R., Laeven, L., & Levine, R. (2016). Does the geographic expansion of banks reduce risk? *Journal of Financial Economics*, *120*, 346–362.

Gude, A., Álvarez, I. C., & Orea, L. (2018). Heterogeneous spillovers among spatial provinces: A generalized spatial stochastic frontier model. *Journal of Productivity Analysis*, *50*, 155–173.

Hirtle, B. (2007). The impact of network size on bank branch performance. *Journal of Banking and Finance*, *31*, 3782–3805.

Horrace, W. C., Parmeter, C. F., & Wright, I. A. (2019). *Probability statements for stochastic frontier models for with spatial errors*: Mimeo.

Jin, F., & Lee, L.-F. (2020). Asymptotic properties of a spatial autoregressive stochastic frontier model. *Journal of Spatial Econometrics*, *1*, 2.

Koetter, M., Kolari, J. W., & Spoerdijk, 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.

Kutlu, L., Tran, K., & 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., & Levine, R. (2007). Is there a diversification discount in financial conglomerates? *Journal of Financial Economics*, *85*, 331–367.

LeSage, J., & Pace, R. K. (2009). *Introduction to Spatial Econometrics*. Boca Raton, Florida: CRC Press,Taylor and Francis Group.

Levine, R., Lin, C., & Xie, W. (2021). Geographic diversification and banks funding costs. *Management Science*, *67*, 2657–2678.

Li, L., Strahan, P. E., & Zhang, S. (2020). Banks as lenders of first resort: Evidence from the COVID-19 crisis. *Review of Corporate Finance Studies*, *9*, 472–500.

Nickell, S. (1981). Biases in dynamic models with fixed effects. *Econometrica*, *49*, 1417–1426.

Orea, L., & Álvarez, I. C. (2019). A new stochastic frontier model with cross-sectional effects in both noise and inefficiency terms. *Journal of Econometrics*, *213*, 556–577.

Orea, L., Álvarez, I. C., & Jamasb, T. (2018). A spatial stochastic frontier model with omitted variables: Electricity distribution in Norway. *Energy Journal*, *39*, 93–116.

Sealey, C., & Lindley, J. T. (1977). Inputs, outputs and a theory of production and cost at depository financial institutions. *Journal of Finance*, *32*, 1251–1266.

Stiroh, K. J. (2004). Diversification in banking: Is noninterest income the answer? *Journal of Money, Credit and Banking*, *36*, 853–882.

Tsionas, E. G., & Michaelides, P. G. (2016). A spatial stochastic frontier model with spillovers: Evidence for Italian regions. *Scottish Journal of Political Economy*, *63*, 243–257.

Wheelock, D. C., & Wilson, P. W. (2001). New evidence on returns to scale and product mix among U.S. commercial banks. *Journal of Monetary Economics*, *47*, 653–674.

Wheelock, D. C., & Wilson, P. W. (2012). Do large banks have lower costs? New estimates of returns to scale for U.S. banks. *Journal of Money, Credit and Banking*, *44*, 171–199.

Wheelock, D. C., & Wilson, P. W. (2018). The evolution of scale economies in USbanking. *Journal of Applied Econometrics*, *33*, 16–28.

Yu, J., de Jong, R., & Lee, L.-F. (2008). Quasi-maximum likelihood estimators for spatial dynamic panel data with fixed effects when both n and T are large. *Journal of Econometrics*, *146*, 118–134.

Zamore, S., Beisland, L. A., & Mersland, R. (2019). Geographic diversification and credit risk in microfinance. *Journal of Banking and Finance*, *109*, 105665.

How to cite this article: Glass, A. J., & Kenjegalieva, K. (2023). Dynamic returns to scale and geography in U. S. banking. *Papers in Regional Science*, 1–33. https://doi.org/10.1111/pirs.12713