Received: 8 January 2022 | Revised: 7 November 2022 | Accepted: 17 December 2022

DOI: 10.1002/ijfe.2776

RESEARCH ARTICLE

WILEY

Returns to scale, spillovers and persistence: A network perspective of U.S. bank size

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

[1]Sheffield University Management School, Sheffield, 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 LE11 3TU, UK,
Email: k.a.kenjegalieva@lboro.ac.uk

Abstract

The methods for ray-scale economies (RSE) and expansion-path scale economies (EPSE) are extended to the dynamic spatial setting. We apply these methods to large U.S. banks using dynamic spatial cost and revenue models and key findings include the following. First, accounting for spillovers and dynamics strengthens the case for EPSE over RSE. Second, own, spillover and total EPSE are very persistent in future periods. Third, the EPSE suggest that an appropriate regulatory size cap would shift one systemically important bank to its contemporaneous optimal scale. However, the EPSE suggest that this would be a sub-optimal dynamic scale in future periods.

KEYWORDS

cost and revenue functions, dynamic spatial modelling, internal and external economies, too-big-to-fail banks

1 | INTRODUCTION

An integral contributing factor to the 2008 financial crisis was the excessive risk taking by very large U.S. banks. This risk taking was promoted by the "too-big-to-fail" (TBTF) status of these banks which was a focus of the 2010 Dodd-Frank regulatory reforms. These reforms took steps to prevent 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 the Dodd-Frank reforms would not prevent TBTF banks and the reforms should have involved taking further preventative action by imposing size caps on the largest U.S. banks. This view is particularly interesting in light of the observation by Wheelock and Wilson (2018) that there were non-negligible post-crisis increases in the sizes of certain very large U.S. banks.

When considering the merits of such size caps, measuring the returns to scale of large U.S. banks is important.[i] This is because size caps will likely lead to a change in the magnitude of these returns and also possibly a change in their classification (increasing / constant / decreasing). The stakeholders which such changes in returns to scale will impact include the relevant particular bank and its employees, society as a whole, the bank's shareholders and bank regulators and antitrust policymakers. Specifically, a change in cost-oriented returns to scale has implications for how efficiently a bank uses society's resources to provide its services (Stern & Feldman, 2009). Additionally, changes in revenue and profit oriented returns to scale will be of concern to a bank's shareholders, as well as to regulators and antitrust policymakers interested in how such changes may impact

industry consolidation (Wheelock & Wilson, 2018). A further important issue is whether the impact of size caps on cost, revenue and profit oriented returns to scale are consistent with one another.[ii] Thus far we have in mind the usual returns to scale that are internal to a bank, but size caps can also influence how external returns to scale impact a bank and other banks and are key component of this paper. External scale economies are the returns to a bank and other banks from the way in which a country's banking industry is organized. External economies are not therefore as tightly defined as internal returns to scale. To address this in one sense we follow Glass, Kenjegaliev, and Kenjegalieva (2020) by using the structure of a spatial empirical model to specify the form of the external returns, and in two other key respects, which we turn to next, our contributions extend their work and thus fill a gap in the literature.

The gap in the literature we identify and the two contributions we make to fill this gap are as follows. The *first contribution* of this paper to the literature is to extend the methods for static external returns to scale in Glass, Kenjegaliev, and Kenjegalieva (2020) to measure the dynamic persistence in future in-sample periods of: (i) returns to scale that are internal to a bank; and (ii) external returns to scale. These new methods enable us to assess whether the classifications of contemporaneous internal and external returns to scale are consistent with the classifications of the dynamic internal and external returns in future periods. These measures of the persistence of internal and external returns can be used to assess the dynamic optimality of, first, a change in the size of a bank and, second and relatedly, some stakeholder policy decisions of bank regulators, such as the decision not to impose size caps. Compared to the familiar static (i.e., contemporaneous) internal returns to scale that are widely reported in the banking literature (e.g., Hughes & Mester, 2013; Wheelock & Wilson, 2001, 2012, 2018), the external and dynamic features of our approach represent a different line of inquiry.

The focus in Glass, Kenjegaliev, and Kenjegalieva (2020) is the presentation of their methods for static internal and external returns to scale and they only provide a small and general demonstration of these methods for the costs of large and medium-sized U.S. banks (1998–2015). In light of this, the *second contribution* of our paper is to carry out a more detailed, policy focused empirical analysis that directly relates to size caps by applying the methods we introduce for dynamic internal and external returns to scale to the costs and revenues of key large U.S. banks (1998–2019); namely, 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). G-SIBs and the banks included in the CCAR are groups where a constituent bank has the capability to influence the global banking system and the domestic banking industry, respectively. Moreover, by considering contemporaneous and dynamic cost and revenue oriented internal and external returns to scale, we are able to analyse whether the implications for bank size from these two sets of findings are consistent.[iii]

We calculate the internal and external returns to scale from a dynamic spatial Durbin model (SDM). The reduced form of the SDM yields what are referred to in the spatial literature as contemporaneous and dynamic direct, indirect and total elasticities (LeSage & Pace, 2009), which we collectively refer to as spatial elasticities and use to calculate the corresponding returns to scale measures. Contemporaneous and dynamic direct elasticities measure the effect of a change in a bank's own independent variable in the current period on its dependent variable in current and future periods, and are used to compute the contemporaneous and dynamic direct (i.e., internal) returns to scale. Also associated with the change in the bank's own independent variable are contemporaneous and dynamic indirect elasticities which measure the spill-in/spill-out to and from the bank in current and future periods, and are used to compute the contemporaneous and dynamic indirect (i.e., external) returns to scale. Summing the direct and indirect spill-in / spill-out elasticities yields two total elasticities which are used to compute the two total returns to scale. The total elasticities represent two network perspectives: (i) the direct impact plus the indirect *spill-in* to a bank is what the bank would focus on and would also be of interest to bank regulators and antitrust policymakers; and (ii) the direct impact plus the indirect *spill-out* from the bank is a network perspective that bank regulators and antitrust policymakers may also find informative, as this has implications for all the banks in the sample. Given both these network perspectives would be of interest to bank regulators, if the two total returns to scale measures are at odds with one another as they indicate that a size cap on a large bank would leave the bank, for example, below and above its optimal size, the policy decision on the imposition of the cap would involve the regulator prioritizing the interests of the bank over the other banks in the industry, or vice-versa.[iv] With regard to some terminology from hereon, when referring to the contemporaneous and dynamic returns to scale we can use internal and direct, and external and indirect, interchangeably. In the parlance of the spatial econometrics literature we use the direct and indirect labels when referring to these returns.

Specifically, we extend the methods for the standard own ray-scale economies (RSE) and expansion-path scale economies (EPSE) to the dynamic spatial setting. The above second contribution of this paper from the empirical analysis of the contemporaneous and dynamic spatial

RSE and EPSE for large U.S. banks involves addressing the following *three research questions* (RQ1–RQ3). During the course of the paper we revisit RQ1–RQ3 in detail.

> **RQ1.** Are there significant contemporaneous and dynamic geographical interdependencies between the costs (revenues) of large U.S. banks, and if so, what are the signs of these interdependencies?
>
> **RQ2.** If corresponding estimates ofRSE and EPSE are at odds with one another, which of the two might be preferred, and how may accounting for spatial interdependencies impact this preference?
>
> **RQ3.** Can a contemporaneous spatial returns to scale measure point to an optimal bank size that is inconsistent with that from the corresponding dynamic measure for one of the next few in-sample periods? If so, what are the recommendations in such a situation?

To provide some initial insights, we summarize the three empirical findings for RQ1–RQ3. For RQ1, significant coefficients on the contemporaneous spatial lag of the cost (revenue) dependent variable and its time lag point to contemporaneous and dynamic geographical interdependencies between the costs (revenues) of large U.S. banks. The contemporaneous and dynamic interdependencies of banks' costs and the contemporaneous interdependency of banks' revenues are all positive, which is consistent with the banks' observations for each of these variables being impacted by common geographical economic phenomena. The dynamic interdependency of banks' revenues is negative, which is consistent with there being a time lag before the impacts of spatial competition take effect.

For RQ2, one can interpret a non-spatial or spatial EPSE measure as more suitable than the corresponding RSE, which is because for banks' costs (revenues) the EPSE measure does not assume that a bank lies on a radial ray in the output (input) space. We find that when we account for spillovers between bank networks, the case for EPSE (which we find to be very persistent in future in-sample periods) over RSE is stronger, and for this reason we focus on spatial EPSE.

For RQ3, to demonstrate the impact of the aforementioned first network perspective of the total EPSE, which is what a bank would focus on, we consider the results for a particular G-SIB, BNY Mellon. From this first perspective we observe a noticeable difference between BNY Mellon's cost oriented contemporaneous and dynamic total EPSE. Such a difference between these contemporaneous and dynamic measures suggests that a bank will be faced with a situation whereby, if it acts on its contemporaneous total EPSE by (not) changing its size in the current period, this can lead to sub-optimal dynamic returns in the following periods. When there is such an inconsistency between these contemporaneous and dynamic total EPSE, we suggest that a bank should aim to optimize its contemporaneous and dynamic total returns over the time frame of its future plans, which would involve some returns being sub-optimal for particular periods within this time frame.

Additionally for RQ3, to demonstrate the impact of the second network perspective of the total EPSE, we consider the results for another G-SIB, JPMorgan Chase. Its cost oriented contemporaneous total EPSE suggests that an appropriate size cap would be consistent with the bank operating at its minimum efficient scale. However, from this second perspective, its cost oriented dynamic total EPSE measures suggest that a size cap would lead to a sub-optimal dynamic scale in the following periods. This and similar other conclusions suggest that dynamic spatial measures can provide additional insights for the stakeholder policy decisions of bank regulators and anti-trust policymakers.

The remainder of this paper is organized as follows. The background in Section 2 motivates our focus on large U.S. banks. To show how we build on the current literature, in Section 3 we set out the standard internal returns to scale theory and via a review of the empirical evidence on returns to scale for U.S. banks we highlight the key features of this literature. In the research design in Section 4, which has four parts, we provide detailed explanations of the three research questions we address. The first part of the research design provides a general presentation of the structural form of the dynamic SDM we estimate; the second shows how the reduced form of this model is used to compute the contemporaneous and dynamic spatial elasticities; the third relates the general models from the first and second parts to the spatial cost and revenue models we consider; and the fourth uses the reduced forms of the spatial cost and revenue models to set out the methods for the contemporaneous and dynamic spatial returns to scale measures. In Section 5, using recent panel data, we present the empirical analysis of the spatial returns to scale measures for large U.S. banks. Section 6 concludes by summarizing the contributions of the returns to scale methods and the practical industry relevance of the empirical findings.

2 | BACKGROUND

In the previous section we touched on how Fisher and Rosenblum (2012), for example, argue that the Dodd-Frank

regulatory reforms would not prevent TBTF U.S. banks. They are therefore of the opinion that these reforms should have involved further preventative action by imposing size caps on the largest banks. Consistent with Fisher and Rosenblum's predictions, we have seen that the Dodd-Frank regulatory reforms did not prevent non-negligible post−2008 increases in the sizes of many large banks. To illustrate, Wheelock and Wilson (2018) note that at the end of 2006, the largest U.S. bank holding company, Citigroup, had total consolidated assets of $1.9 trillion, with a further two (Bank of America and JPMorgan Chase) having assets in excess of $1 trillion. In comparison, at the end of 2015, they also report that the largest holding company was JPMorgan Chase with assets of $2.35 trillion, with three others having assets of more than $1.7 trillion. One interpretation of these size increases is that they make a stronger case for size caps on the largest banks as a means of more concerted regulatory action against these banks being TBTF. One key reason why we focus on very large U.S. banks is to examine whether these banks are characterized by cost and revenue diseconomies of scale, as this would be consistent with this stronger case. If, on the other hand, there is an inconsistency, then from the cost and / or revenue perspective(s) a size cap will move a very large bank further below its optimal size(s). From the cost perspective this would not be in society's interests as it would involve the bank using society's resources more inefficiently to provide its services, while from the revenue perspective a more sub-optimal size would not be in the interests of bank shareholders. Turning next to further reasons why we consider these banks and others that collectively represent our sample of large U.S. banks.

We carry out a policy focused empirical analysis that directly relates to size caps by applying the methods for contemporaneous and dynamic spatial returns to scale to the key largest banks in our sample; namely, global systemically important banks (G-SIBs) (Financial Stability Board, 2019) and banks included in the *Comprehensive Capital Analysis and Review* (CCAR) (Federal Reserve Board, 2019). We consider these banks because the definitions of a G-SIB and CCAR bank are consistent with these banks being TBTF. This is because G-SIBs and CCAR banks are groups where a constituent bank has the capability to influence the global banking system and the domestic banking industry, respectively. The contemporaneous and dynamic spatial returns to scale are well-suited to G-SIBs and CCAR banks, as the capabilities of these banks to impact the global banking system and the domestic banking industry will be determined by, among other things, how the geographical interconnectedness of the bank impacts how its size affects (and is affected by) the sizes of other banks. This highlights the importance of accounting for indirect returns to scale spill-ins and spill-outs and assessing how they affect the optimal sizes of G-SIBs and CCAR banks. We assess this by comparing a bank's contemporaneous and dynamic direct-own EPSE with its corresponding two total EPSE.

We also consider large U.S. banks because their geographical interconnectedness, which we measure using the degree of overlap between banks' branch networks, is more marked than it is for smaller banks. This is because large banks have bigger branch networks which means that there will be more cases where banks operate in the same market, where a banking market is taken to be a metropolitan statistical area (MSA) or non-MSA county (Hirtle, 2007). Given the branch network of each large bank overlaps with a sufficient number of networks of the other banks, there is no lack of geographical interconnectedness between the banks in our sample, and so spatial methods are therefore well suited to modelling the spatial interactions between these banks. In these methods the geographical interconnectedness between the banks is accounted for using what is referred to as the spatial weights matrix. See Section 5.1 for details on how we go about the a priori construction of this matrix for the empirical analysis. Finally, and crucially, U.S. banks are the subject of the empirical analysis due to the rich data that is available on the zip codes of the locations of all the banks' branches. Using this branch location data we specify the geographical linkages between the banks in the spatial weights matrix.

3 | REVIEW OF RETURNS TO SCALE THEORY AND BANKING EVIDENCE

The vast majority of the literature on bank returns to scale adopt the traditional approach and estimate RSE and EPSE that are strictly internal to a bank. Of these two measures internal RSE is by far the most commonly estimated. For the case of a bank with two outputs or inputs in Figure 1, internal RSE relates to an equiproportional change in both of the bank's outputs or inputs along the radial ray 0A. From an estimated cost function, a rise in cost that is less than, equal to, or greater than an increase in outputs along 0A corresponds to internal RSE that is $<$, $=$, or >1, and represents increasing, constant or decreasing returns to scale, respectively. Alternatively, from an estimated revenue (alternative profit) function, a rise in revenue (profit) that is less than, equal to, or greater than an increase in inputs (outputs) along 0A corresponds to internal RSE that is $<$, $=$, or >1, and represents decreasing, constant, or increasing returns to scale. The implications of increasing, constant and decreasing

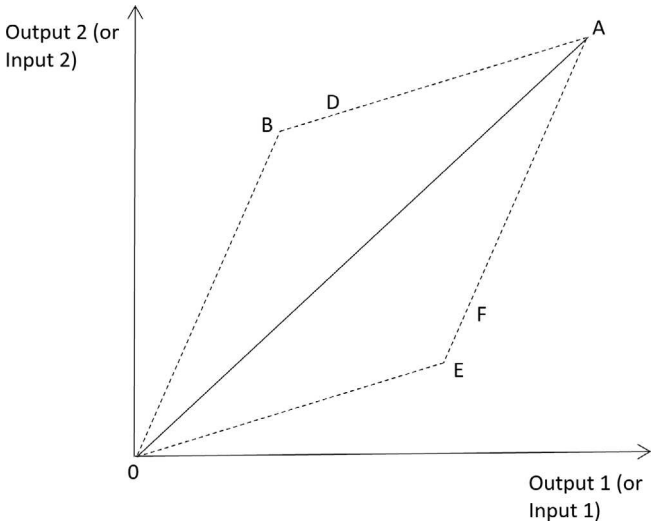

FIGURE 1 Output and input levels for internal ray-scale and internal expansion-path scale economies

internal RSE is that from a cost, revenue or profit perspective, the bank is smaller than, equal to, or larger than its optimal size. Berger et al. (1987) recognized that a bank may not lie on a radial ray and proposed internal EPSE to account for this. The classification of internal EPSE (increasing / constant / decreasing) from an estimated cost, revenue or alternative profit function and the associated bank size implications of the internal EPSE are as above for the internal RSE. However, in contrast to internal RSE, internal EPSE corresponds to incremental changes in outputs or inputs along a non-radial ray, such as BD or EF in Figure 1, where these rays represent portions of the bank's output or input expansion-paths 0BA or 0EA.

Moving on to the empirical research on returns to scale of U.S. banks, where in our following coverage of this research we draw a distinction between the two strands of the literature. In particular, we highlight the key features of each strand by drawing together the major similarities and differences between studies. The studies in first strand, which is by far the largest of the two strands, estimate the commonly reported standard internal returns to scale, where the vast majority of these studies estimate internal RSE. With regard to the empirical estimates from this first strand, a key feature of this literature is that the classifications of internal returns to scale (increasing / constant / decreasing) are in many cases mixed. More specifically, and as we indicate in the following discussion of this literature, the evidence on these returns is not robust across different study periods, bank sizes, functional forms and estimation methods. For instance, Feng and Zhang (2014) and Restrepo-Tobón and Kumbhakar (2015) present evidence of decreasing internal RSE in the operations of some large U.S. banks. Both these studies use a sample period that ends in 2010 and a distance function approach, with the former study using a Bayesian estimation procedure and the latter a non-parametric estimator. Likewise, using a sample for 1986, Noulas et al. (1990) estimate a parametric translog cost function and observe decreasing internal RSE for large U.S. banks with assets between $3−$6 billion.

Of the vast majority of studies in the first strand that estimate internal RSE for U.S. banks, there are noticeably more studies that report increasing and/or constant internal RSE than there are studies that find decreasing internal RSE. The studies that report increasing and / or constant internal RSE use a wide range of approaches, such as Bayesian estimation (Feng & Serletis, 2010), the Almost Ideal Demand System (Hughes & Mester, 2013) and non-parametric methods (Wheelock & Wilson, 2012, 2018), and in some cases report quite substantial increasing returns to scale. Kovner et al. (2014) is an interesting further study that reports increasing returns to scale for a particular expenditure category. Specifically, they find that a 10% increase in assets can lead to a 0.3%–0.6% decline in non-interest expenses. In contrast, when total cost of U.S. banks is considered rather a particular expenditure category, Feng and Zhang (2014) find that there is no clear pattern in the relationship between asset size and the classification of returns to scale. Some studies also report a mixture of constant and increasing returns to scale across estimation methods (Wheelock & Wilson, 2001) and across cost, revenue and alternative profit functions (Wheelock & Wilson, 2018). The above findings on increasing and / or constant internal RSE remain broadly unchanged for internal EPSE estimates for U.S. banks (Wheelock & Wilson, 2018).

As is often the case in panel or cross-sectional banking studies, studies that estimate internal RSE and EPSE for U.S. banks consider the typical setting where the model errors in each cross-section are taken to be spatially independent. With this approach, if the errors in each cross-section are not spatially independent, this will at the very least invalidate the statistical inference (i.e., the standard errors), and also potentially involve a model where there is an omitted variable bias because of the exclusion of spatially lagged independent variables and / or the spatial lag of the dependent variable. In studies that estimate internal RSE and EPSE using a sample that includes large U.S. banks one may expect there to be such spatial autocorrelation. This is because these large banks are often (but not always) among the banks with the largest branch networks and so it follows that there is greater geographical overlap between these networks. Hence there are more cases where large banks operate in the same markets, which is consistent with their being

spatial autocorrelation for two reasons. First, there will be more geographical competition between these banks and, second, the greater the geographical overlap between banks' branch networks, the more they will be exposed to common geographical economic phenomena, such as market growth and headline changes in regional, state and city economies.

Such spatial autocorrelation is accounted for in the much smaller and emerging second strand of the literature on returns to scale of U.S. banks (Glass, Kenjegalieva, & Douch, 2020; Glass & Kenjegalieva, 2019; Glass, Kenjegaliev, & Kenjegalieva, 2020; Glass, Kenjegalieva, & Weyman-Jones, 2020). Consistent with there being spatial autocorrelation among U.S. banks these studies report clear evidence of significant spatially lagged variables. The significance of the spatial lag of the dependent variable in the relevant model, such as a spatial cost model, is what allows one to proceed and compute external returns to scale, for example, associated with a change in a bank's outputs are: (a) cost returns that spill-in to a bank from other banks; and (b) returns that spill-out from the bank to others. Of these papers Glass, Kenjegaliev, and Kenjegalieva (2020) focus on returns to scale spillovers and is the paper in the second strand of the literature that we concentrate on, while Glass and Kenjegalieva (2019) and Glass, Kenjegalieva, and Douch (2020) compute the growth of RSE spillovers as part of their focus on spatial total factor productivity growth decompositions.

Glass, Kenjegaliev, and Kenjegalieva (2020) consider the usual contemporaneous internal RSE and EPSE, as well as contemporaneous external spillover RSE and EPSE and the corresponding contemporaneous total measures that incorporate the internal and external estimates. Using panel data for large and medium-sized U.S. banks (1998–2015), they estimate a static spatial cost model and, for the full sample of U.S. banks and subsamples of large banks and medium-sized banks, report average contemporaneous internal RSE and EPSE that are, in general, not significantly different from 1 (although in some cases are less than 1 in magnitude). These findings provide support for some of the results in a number of the aforementioned studies (Hughes & Mester, 2013; Wheelock & Wilson, 2012, 2018) and suggest that when a bank is considered in isolation (i.e., spatial interactions between banks are overlooked), on average, bank size is not statistically sub-optimal.

Interestingly, Glass, Kenjegaliev, and Kenjegalieva (2020) also find for their full sample, large banks and medium-sized banks that the average contemporaneous external RSE are not significantly different from zero, which is due to the offsetting signs of the spatially lagged independent variables and the spatial lag of dependent variable, while the corresponding average contemporaneous external EPSE are not significantly different from 1. The upshot is that for the full sample, large banks and medium-sized banks they report both total RSE and EPSE that are, in general, not significantly different from 1. Their headline conclusion from these results is that when spatial interactions between banks are accounted for, on average, bank size is not statistically sub-optimal. Given this conclusion is based on total RSE and EPSE that are consistent with one another, which importantly and rather unusually is due to very different external RSE and EPSE that are likely due to a particular rare feature of the data in Glass, Kenjegaliev, and Kenjegalieva (2020), in our paper we investigate three issues when the external RSE and EPSE differ and we observe the more likely finding in this situation of different total RSE and EPSE. The first issue is whether the bank size implications of contemporaneous total RSE and EPSE remain the same when we estimate spatial cost and revenue functions using updated data for only large U.S. banks. Second, if there is an inconsistency between corresponding contemporaneous (dynamic) total RSE and EPSE, which of the two measures do we favour. Third, if there is an inconsistency between the contemporaneous estimate of our preferred total returns to scale measure and the corresponding dynamic estimate for one of the next few in-sample periods, how is the inconsistency addressed to arrive at a clear and unequivocal policy recommendation on what is the optimal size of a particular bank.

4 | RESEARCH DESIGN

4.1 | Dynamic spatial model with fixed effects

The structural form of the dynamic SDM with fixed effects for balanced panel data that we estimate is

$$\mathbf{y}_t = \alpha\boldsymbol{\iota} + \boldsymbol{\beta}'\mathbf{X}_t + \boldsymbol{\gamma}'\mathbf{W}\mathbf{X}_t + \boldsymbol{\eta}'\mathbf{W}\mathbf{X}_{t-1} + \delta\mathbf{W}\mathbf{y}_t + \lambda\mathbf{W}\mathbf{y}_{t-1} + \boldsymbol{\zeta} + \boldsymbol{\varepsilon}_t. \quad (1)$$

The balanced panel data comprises observations for T periods (indexed $t \in 1, ..., T$) and N banks (indexed i, $j \in 1, ..., N \forall\, i \neq j$).[v] $\mathbf{y}_t$ is the N-dimensional stacked vector of logged cost or revenue observations, $\mathbf{X}_t$ is the stacked $N \times K$ matrix of logged observations of the non-spatial regressors (indexed $k \in 1, ..., K$), and $\boldsymbol{\beta}'$ is the associated vector of parameters. $\boldsymbol{\iota}$ and $\boldsymbol{\zeta}$ are N-dimensional vectors of ones and fixed effects, respectively, α is the common intercept and $\boldsymbol{\varepsilon}_t$ is the N-dimensional stacked vector of idiosyncratic disturbances. To account for unobserved heterogeneity, rather than use, for example,

random effects, we use fixed effects so that there can be correlation between the effects and the regressors.

$\mathbf{W}$ is the fixed $N \times N$ spatial weights matrix comprising the non-negative weights w_{ij}. $\mathbf{W}$ is specified a priori and represents: (i) the spatial arrangement of the N banks in the cross-sections; and (ii) the strength of the spatial interaction between these banks. Typically, $\mathbf{W}$ is exogenous, which is also an assumption we make about $\mathbf{W}$ in Equation 1. In line with this exogeneity, a measure of geographical proximity is frequently used to specify the spatial weights, which is the approach we adopt in the empirical analysis using a novel measure of geographical interconnectedness. Since a bank cannot be linked to itself, all the elements on the main diagonal of $\mathbf{W}$ are set to zero. $\mathbf{W}\mathbf{y}_t$ is the contemporaneous spatial lag of the dependent variable, otherwise known as the SAR variable. Our model is dynamic because, for reasons we discuss further in this section, rather than include a time lag of the dependent variable, we include $\mathbf{W}\mathbf{y}_{t-1}$. This is because it is entirely reasonable for it to take some time for spillovers to occur. The SAR parameters $\{\delta, \lambda\} \in (1/d_{\min}, 1/d_{\max})$, where $d_{\min}$ and $d_{\max}$ are the most negative and positive real characteristic roots of $\mathbf{W}$. Note that $\mathbf{W}$ denotes a normalized specification of the spatial weights matrix, where the normalization we use in the empirical analysis gives $d_{\max} = 1$. For details of this normalization see Section 5.1 in the empirical analysis.

In a number of respects we follow influential non-spatial U.S. banking returns to scale studies (Wheelock & Wilson, 2012, 2018) by estimating a theoretical functional form of the technology. We therefore omit $\mathbf{y}_{t-1}$ as it does not form part of the theoretical function. For the same reason, we also omit non-spatial variables that shift the frontier. In our spatial setting this has the benefit of parsimony because given the translog functional form we use and our focus on modelling spillovers, we have quite a number of spatial variables that shift the frontier. As noted by Orea et al. (2018), modelling spatial dependence also represents a way to mitigate omitted variables and the resulting endogeneity. In this respect we are comprehensive as we include a range of spatial variables that shift the frontier. These spatial variables fall into two categories. (i) $\mathbf{W}\mathbf{X}_t$ and $\mathbf{W}\mathbf{X}_{t-1}$ ($N \times K$ matrices of spatially lagged observations, where $\boldsymbol{\gamma}'$ and $\boldsymbol{\eta}'$ are the associated vectors of parameters) are exogenous local spatial regressors that account for only spatial interaction between a bank and its 1st order neighbours. Some dynamic SDMs in the literature include $\mathbf{W}\mathbf{X}_t$, but not its one period time lag (e.g., Ciccarelli & Elhorst, 2018 and LeSage & Sheng, 2014). This is likely for parsimony, whereas we include $\mathbf{W}\mathbf{X}_{t-1}$ to be consistent with the rationale for the inclusion of $\mathbf{W}\mathbf{y}_{t-1}$: namely, both $\mathbf{W}\mathbf{X}_{t-1}$ and $\mathbf{W}\mathbf{y}_{t-1}$ capture spillovers that take some time to occur. (ii) By including $\mathbf{W}\mathbf{y}_t$ and $\mathbf{W}\mathbf{y}_{t-1}$ the reduced form of Equation 1 (see Equation 2) accounts for contemporaneous and dynamic global spatial interactions: namely, contemporaneous and dynamic spatial interactions between a bank and its 1st order, 2nd order, 3rd order, etc. neighbouring banks.

Using the fitted dynamic spatial cost and revenue models, we address the following first research question:

Are there significant contemporaneous and dynamic geographical interdependencies between the costs (revenues) of large U.S. banks, and if so, what are the signs of these interdependencies

Tackling RQ1 will inform whether and how the geographical interconnectedness between banks impacts their costs and revenues. More specifically, in RQ1 contemporaneous interdependency between banks' costs (revenues) relates to the coefficient on the contemporaneous SAR variable, while the corresponding dynamic interdependency is captured by the coefficient on the time lag of this variable. Thus far, only contemporaneous interdependency between U.S. banks' costs (revenues) has been analysed using static spatial models. Over the period 1992–2015 for large U.S. banks (Glass & Kenjegalieva, 2019), and 1998–2015 for large and medium-sized U.S. banks (Glass, Kenjegaliev, & Kenjegalieva, 2020), the significant non-negligible SAR parameters point to positive contemporaneous interdependency between banks' costs. For 1998–2015 and large and medium-sized U.S. banks (Glass, Kenjegalieva, & Douch, 2020), the significant non-negligible SAR parameters indicate positive contemporaneous interdependency between banks' revenues.[vi] Although we use more recent data for large U.S. banks (1998–2019), we expect that the contemporaneous SAR parameters in our dynamic spatial models would also be significant, non-negligible and point to positive contemporaneous interdependencies between banks' costs and revenues. However, compared to the aforementioned results from static spatial models, the inclusion of a time lag of the SAR variable may well impact the magnitude of a positive contemporaneous SAR parameter and could even lead to a change in its sign. There is no study that has estimated a dynamic spatial model for banks that we can draw to indicate what results we might expect for the dynamic SAR parameters. That said, one might think that positive dynamic SAR parameters are more likely as negative spatial interdependency is less common in the spatial literature. Positive spatial interdependency is associated with units which face common geographical economic phenomena; while negative spatial interdependency is attributed to the effects of competition (e.g., Boarnet & Glazer, 2002 and Garrett & Marsh, 2002). For example, a decrease in a firm's revenue is the spatial competitive effect of a rise in the revenues of its neighbouring firms.

To estimate Equation (1) we use the quasi-maximum likelihood (QML) approach in Yu et al. (2008). ML assumes normally distributed errors, whereas QML is less restrictive as it involves no such assumption. The estimation of our model has three further features. First, as is standard for fixed effects models, we use the within transformation to circumvent the incidental parameter problem. Second, the estimator corrects for the biases from the fixed effects (Nickell, 1981). Third, the concentrated log-likelihood function includes $T\log|\mathbf{I}_N-\delta\mathbf{W}|$, which is the scaled logged determinant of the Jacobian of the transformation from $\boldsymbol{\varepsilon}_t^*$ to $\mathbf{y}_t^*$. $\mathbf{I}_N$ is the N-dimensional identity matrix and a * denotes the demeaned transformations of $\boldsymbol{\varepsilon}_t$ and $\mathbf{y}_t$. As is standard in spatial econometrics, the transformation from $\boldsymbol{\varepsilon}_t^*$ to $\mathbf{y}_t^*$ accounts for the endogeneity of the contemporaneous SAR variable and also the fact that $\boldsymbol{\varepsilon}_t$ is not observed (Anselin, 1988; Elhorst, 2009).[vii,viii]

4.2 | Contemporaneous and dynamic elasticities to measure global spillovers

The coefficients on $\mathbf{X}_t$ and $\mathbf{X}_{t-1}$ in the structural form in Equation (1) are own elasticities. From the same equation, the coefficients on $\mathbf{WX}_t$ and $\mathbf{WX}_{t-1}$ are elasticities that represent the local spillovers to a bank from marginal changes in these weighted contemporaneous and dynamic independent variables of its 1st order neighbours. These elasticities do not therefore represent the global spillovers to a bank from marginal changes in these variables of its higher order (as well as its 1st order) neighbours. This is because elasticities that account for global spillovers, which are referred to as direct, indirect and total impacts, are a function of the SAR parameter(s). Using the fitted parameters from the structural form of our model we compute the contemporaneous and dynamic direct, indirect and total elasticities of the $\mathbf{X}_t$ variables.

A contemporaneous direct elasticity is interpreted in the same way as a contemporaneous own elasticity from a non-spatial model, although the direct elasticity takes into account feedback. This feedback is the contemporaneous effect of a change in an independent variable for a particular bank which partially reverberates back to the same bank's dependent variable through its effect on the dependent variables of the other banks in the sample. A contemporaneous indirect elasticity for an individual bank can be calculated in two ways: (i) average change in the dependent variable of all the other banks in the sample due to a change in an independent variable for one bank that is, a spill-out from one bank; or (ii) average change in the dependent variable for one bank due to a change in an independent variable of all the other banks in the sample that is, a spill-in to one bank. To facilitate interpretation it is common to report an average indirect elasticity across the N banks. As will become evident from the formal presentation below, averaging (i) or (ii) across all the N banks yields the same value. In the empirical analysis, since the average indirect (i) and (ii) elasticities across the N banks are of the same magnitude, we obtain the same average value for the contemporaneous indirect spill-in and spill-out returns to scale.

Summing a variable's contemporaneous direct elasticity and indirect elasticity (i) or (ii) gives the corresponding total elasticity. We therefore obtain the same average value for the total elasticity using the average of (i) or (ii) across the N banks. Hence, using the average of (i) or (ii) across the N banks, we obtain the same average value for the two measures of the contemporaneous total returns to scale.

However, for an individual bank (or an average across any subset of the N banks), the two contemporaneous indirect elasticities will be of different magnitudes. As a result, for individual banks there will be an asymmetry between the two contemporaneous total elasticities and between the two contemporaneous indirect (total) returns to scale.

From the spatial model we also obtain average measures across the N banks of the corresponding dynamic elasticities and returns to scale (direct, symmetric indirect spill-out and spill-in, and the resulting two symmetric total measures). These dynamic elasticities quantify how marginal changes in the $\mathbf{X}$ variables in period t impact on $\mathbf{y}$ in future in-sample periods. For individual banks, the dynamic indirect and total elasticities are asymmetric and are used to compute asymmetric indirect and total returns to scale.

To compute the contemporaneous and dynamic direct, indirect and total elasticities for the $\mathbf{X}$ variables involves using the reduced form of Equation 1. In particular, to compute these elasticities we make a minor adjustment to the approach in Debarsy et al. (2012) to account for the omission of $\mathbf{y}_{t-1}$ in our model. At the outset, we note that we condition on the vector of observations of $\mathbf{y}$ for the initial period and assume that this period is only subject to spatial dependence. We can therefore write the dependent variable for the whole sample as $\widetilde{\mathbf{Y}}=\left(\mathbf{y}_1',\ldots,\mathbf{y}_T'\right)'$. The reduced form of Equation (1) that is, the data generating process (DGP) of $\widetilde{\mathbf{Y}}$, is therefore as follows.

$$\widetilde{\mathbf{Y}}=\sum_{k=1}^{K}\mathbf{Z}^{-1}\left[\beta_k\mathbf{I}_{NT}+\gamma_k\left(\mathbf{I}_T\bigotimes\mathbf{W}\right)+\eta_k\left(\mathbf{I}_T\bigotimes\mathbf{W}\right)\right]\widetilde{\mathbf{X}}_k+\mathbf{Z}^{-1}(\alpha\boldsymbol{\iota}_{NT}+\mathbf{B}\boldsymbol{\zeta}+\boldsymbol{\varepsilon}),\tag{2}$$

where

$$\mathbf{Z}=\begin{pmatrix}\mathbf{U} & \mathbf{0} & \cdots & & \mathbf{0}\\ \mathbf{V} & \mathbf{U} & & & \mathbf{0}\\ \mathbf{0} & \mathbf{V} & \ddots & & \vdots\\ \vdots & & \ddots & & \mathbf{0}\\ \mathbf{0} & \cdots & & \mathbf{V} & \mathbf{U}\end{pmatrix} \text{ and } \mathbf{Z}^{-1}$$

$$=\begin{pmatrix}\mathbf{U}^{-1} & \mathbf{0} & \cdots & & \mathbf{0}\\ \mathbf{L}_1 & \mathbf{U}^{-1} & & & \mathbf{0}\\ \mathbf{L}_2 & \mathbf{L}_1 & \ddots & & \vdots\\ \vdots & & \ddots & & \mathbf{0}\\ \mathbf{L}_{T-1} & \mathbf{L}_{T-2} & \cdots & \mathbf{L}_1 & \mathbf{U}^{-1}\end{pmatrix}.$$

$\mathbf{B}$ is the $NT \times N$ matrix that assigns each of the fixed effects to the corresponding bank in each period of the sample. $\mathbf{Z}$ denotes the time and space filtered $NT \times NT$ block matrix, where $\mathbf{U}=(\mathbf{I}_N-\delta\mathbf{W})$ and $\mathbf{V}=-\lambda\mathbf{W}\times\mathbf{Z}^{-1}$ denotes the $NT \times NT$ lower triangular block matrix, and the time filtered $\widetilde{\mathbf{X}}_k$ denotes the *kth* column of the $NT \times K$ matrix $\mathbf{X}$. The current period and remaining periods in the dataset (where the latter represent the in-sample future time horizons) in $\mathbf{Z}^{-1}$ are indexed $\phi \in 0, 1, ..., T-2, T-1$ that is, 0 denotes the contemporaneous period and, for any time period in the dataset, the number of future in-sample time periods can range from 1 through to $T-1$. This highlights that the DGP in Equation (2) is in a form that can be used to compute the partial derivative of $\widetilde{\mathbf{Y}}$ for each of the ϕ periods with respect to a marginal change in $\mathbf{x}_{kt}$, where these partial derivatives yield the contemporaneous and dynamic direct, indirect and total elasticities. $\mathbf{L}_\phi=(-1)^\phi(\mathbf{U}^{-1}\mathbf{V})^\phi\mathbf{U}^{-1}$, and so it is clear that to compute these elasticities we only need to compute $\mathbf{U}^{-1}$ and $\mathbf{V}$.

$\partial\widetilde{\mathbf{Y}}_{t+1}/\partial\widetilde{\mathbf{X}}_{kt}'$ is the matrix of one period ahead partial derivative impacts of a permanent marginal change in $\mathbf{x}_k$ at time t.

$$\frac{\partial\widetilde{\mathbf{Y}}_{t+1}}{\partial\widetilde{\mathbf{X}}_{kt}'}=\left(\mathbf{L}_1+\mathbf{U}^{-1}\right)(\beta_k\mathbf{I}_N+\gamma_k\mathbf{W}+\eta_k\mathbf{W}). \quad (3)$$

By a permanent marginal change in $\mathbf{x}_k$ at time t, we mean that the values of this variable increase to a new level and remain there in future periods that is, $\partial\widetilde{\mathbf{X}}_{kt}'=(\mathbf{x}_{kt}+\rho, \mathbf{x}_{kt+1}+\rho, ..., \mathbf{x}_{kT}+\rho)$. The elements on the main diagonal of $\partial\widetilde{\mathbf{Y}}_{t+1}/\partial\widetilde{\mathbf{X}}_{kt}'$ represent the one period ahead direct elasticities for each of the N banks. The sums of the off-diagonal elements along a row (and the corresponding column) of this matrix represent a bank's two asymmetric one period ahead indirect elasticities. They measure the one period ahead spill-in (spill-out) to (from) the bank from (to) all the other banks. The two summations for a bank of its direct elasticity and, in turn, each indirect elasticity yields its two asymmetric one period ahead total elasticities. As we noted above, although for an individual bank (or an average across any subset of the N banks) the two indirect (total) elasticities are asymmetric, averaging the two indirect (total) elasticities across the N banks yields the same value. This is because, across the N banks, the average one period ahead total impact is the average of the row or column sums of $\partial\widetilde{\mathbf{Y}}_{t+1}/\partial\widetilde{\mathbf{X}}_{kt}'$, while the average indirect one period ahead impact is the average of the sums of the off-diagonal elements of these rows or columns.

Let Φ denote the set of periods from the contemporaneous period 0 up to ϕ periods ahead. $\partial\widetilde{\mathbf{Y}}_\phi/\partial\widetilde{\mathbf{X}}_{kt}'$ is the matrix of (cumulative) ϕ periods ahead partial derivative impacts of a permanent marginal change in $\mathbf{x}_k$ at time t. Computing this matrix involves cumulating down the columns of $\mathbf{Z}^{-1}$.

$$\frac{\partial\widetilde{\mathbf{Y}}_\phi}{\partial\widetilde{\mathbf{X}}_{kt}'}=\sum_{\phi\in\Phi}\mathbf{L}_\phi(\beta_k\mathbf{I}_N+\gamma_k\mathbf{W}+\eta_k\mathbf{W}). \quad (4)$$

Recall that calculating the average one period ahead direct, symmetric indirect and symmetric total elasticities across the N banks involves taking averages of the sums of the relevant elements of the $N \times N$ matrix $\partial\widetilde{\mathbf{Y}}_{t+1}/\partial\widetilde{\mathbf{X}}_{kt}'$. In a similar way we calculate the average (cumulative) ϕ periods ahead direct, symmetric indirect and symmetric total elasticities across the N banks. This involves taking averages of the same sums of the elements of the $N \times N$ matrix $\partial\widetilde{\mathbf{Y}}_\phi/\partial\widetilde{\mathbf{X}}_{kt}'$. For individual banks (or an average across any subset of the N banks), the asymmetric measures of the (cumulative) ϕ periods ahead indirect and total elasticities are computed in the same way as the corresponding one period ahead elasticities.

Following the spatial literature (e.g., LeSage & Pace, 2009), statistical inference for the contemporaneous and dynamic direct, indirect and total elasticities is via 1, 000 Monte Carlo simulations.

4.3 | Relating direct, indirect and total elasticities to translog functions

In the empirical analysis we estimate dynamic spatial cost and revenue functions using the model specification in Equation (1). We use the following spatial translog functional forms for the models, which involves, among other things, including as regressors the spatial lag of the translog specification of the technology in period t and its

one period time lag. Note to aid the interpretation we present various components of Equations (5) and (6) in scalar form for example, an individual spatial weight w_{ij}.

$$c_{it} = \alpha + TL(t, \mathbf{q}, \mathbf{s})_{it} + \sum_{j=1}^{N} w_{ij} TL(t, \mathbf{q}, \mathbf{s})_{jt} + \sum_{j=1}^{N} w_{ij} TL(t, \mathbf{q}, \mathbf{s})_{jt-1} + \delta \sum_{j=1}^{N} w_{ij} c_{jt} + \lambda \sum_{j=1}^{N} w_{ij} c_{jt-1} + \zeta_i + \varepsilon_{it}, \tag{5}$$

$$r_{it} = \alpha + \mathrm{TL}(t, \mathbf{m}, \mathbf{p})_{it} + \sum_{j=1}^{N} w_{ij} \mathrm{TL}(t, \mathbf{m}, \mathbf{p})_{jt} + \sum_{j=1}^{N} w_{ij} \mathrm{TL}(t, \mathbf{m}, \mathbf{p})_{jt-1} + \delta \sum_{j=1}^{N} w_{ij} r_{jt} + \lambda \sum_{j=1}^{N} w_{ij} r_{jt-1} + \zeta_i + \varepsilon_{it}. \tag{6}$$

c_{it} and r_{it} are logged total cost and total revenue observations for the ith bank in period t. $\mathbf{q}$, $\mathbf{s}$, $\mathbf{m}$ and $\mathbf{p}$ are logged vectors of observations of the outputs, input prices, inputs and output prices, respectively. Each of the dependent variables in Equations (5) and (6) together with the translog function TL_{it} represent the frontier technology and all the other terms shift the frontier. t, t^2 and interactions between t and the other first order variables form part of TL_{it} and collectively represent a non-linear time trend that measures non-neutral technical change. Everything else is as previously defined. Details of the data that is used in the empirical analysis for the variables in Equations (5) and (6) is provided in Section 5.1.

To compute the estimates of the spatial scale economies for period ϕ, we use five translog equations for different cost and revenue measures: direct; two indirect (spill-in and spill-out); and thus two total. We obtain these five translog cost and revenue equations from the estimates of Equations (5) and (6). In the empirical analysis, for the G-SIBs and the banks included in the CCAR, we compare their direct returns to scale, which are essentially own returns as there is very little feedback in our results, with their two total returns that account for the two different network effects. We therefore in Equation (7) illustrate the form of one of the two total translog revenue equations that incorporates revenue spill-ins to a bank from the other banks in the sample $\left(r^{\mathrm{Tot}}_{\mathrm{In},i\phi}\right)$. The translog equations for the following have a similar form to Equation 7, but for brevity we do not present them: a bank's direct revenue $\left(r^{\mathrm{Dir}}_{i\phi}\right)$; the indirect revenue spill-in to a bank from the other banks $\left(r^{\mathrm{Ind}}_{\mathrm{In,if}}\right)$; the indirect revenue spill-out from a bank to the other banks $\left(r^{\mathrm{Ind}}_{\mathrm{Out},i\phi}\right)$; the other total revenue measure that incorporates revenue spill-outs from a bank $\left(r^{\mathrm{Tot}}_{\mathrm{Out},i\phi}\right)$; and the corresponding five translog cost equations. To illustrate, in the other four translog revenue equations, the independent variables are as in Equation (7), with the subscripts and superscripts of the parameters (ξ and τ) and vectors ($\boldsymbol{\theta}'$; ϱ'; $\boldsymbol{\varphi}'$; $\boldsymbol{\varpi}'$) and matrices ($\boldsymbol{\Upsilon}$; $\boldsymbol{\Omega}$; $\boldsymbol{\Psi}$) of parameters matching those of the dependent variable.

$$r^{\mathrm{Tot}}_{\mathrm{In},i\phi} = \xi^{\mathrm{Tot}}_{\mathrm{In},i\phi} t + \frac{1}{2}\tau^{\mathrm{Tot}}_{\mathrm{In},i\phi} t^2 + \boldsymbol{\theta}^{\mathrm{Tot}'}_{\mathrm{In},i\phi} \mathbf{m}_{it} + \varrho^{\mathrm{Tot}'}_{\mathrm{In},i\phi} \mathbf{p}_{it} + \frac{1}{2}\mathbf{m}'_{it} \boldsymbol{\Upsilon}^{\mathrm{Tot}}_{\mathrm{In},i\phi} \mathbf{m}_{it} + \frac{1}{2}\mathbf{p}'_{it} \boldsymbol{\Omega}^{\mathrm{Tot}}_{\mathrm{In},i\phi} \mathbf{p}_{it} + \mathbf{m}'_{it} \boldsymbol{\Psi}^{\mathrm{Tot}}_{\mathrm{In},i\phi} \mathbf{p}_{it} + \boldsymbol{\varphi}^{\mathrm{Tot}'}_{\mathrm{In},i\phi} \mathbf{m}_{it} t + \varpi^{\mathrm{Tot}'}_{\mathrm{In},i\phi} \mathbf{p}_{it} t. \tag{7}$$

A subscript ϕ is attached to the parameters and the dependent variable in Equation (7) to indicate that a parameter relates to the impact in horizon ϕ of a marginal change in a bank's independent variable in t. When t and ϕ correspond to the same period, which is only when $\phi = 0$, these parameters are contemporaneous. When this is not the case, the parameters in Equation (7) are dynamic. This indicates that a change in a bank's independent variable in t will impact the dependent variable ϕ (in-sample) periods ahead.

There are some differences between Equations (5) and (6) and the direct, indirect and total translog cost and revenue functions. Unlike Equations (5) and (6), the direct, indirect and total functions are not regressions. This is because the dependent variables in the direct, indirect and total equations are not observed and so there are no error terms in these equations. One can use the direct, indirect and total equations to compute these unobserved dependent variables, where, to illustrate, $r^{\mathrm{Tot}}_{\mathrm{In},i\phi} = r^{\mathrm{Dir}}_{i\phi} + r^{\mathrm{Ind}}_{\mathrm{In},i\phi}$ and $r^{\mathrm{Tot}}_{\mathrm{Out},i\phi} = r^{\mathrm{Dir}}_{i\phi} + r^{\mathrm{Ind}}_{\mathrm{Out},i\phi}$. In addition, in contrast to Equations (5) and (6), none of the observations in the direct, indirect and total equations are premultiplied by the sum of the spatial weights. This is because the effect of the spatial weights is incorporated within the direct, indirect and total parameter estimates.

4.4 | Contemporaneous and dynamic spatial returns to scale

Using the five (direct, indirect spill-in and spill-out, and hence two total) cost and revenue translog functions for horizon ϕ (e.g., Equation 7 is one of the two total revenue

functions), we calculate the corresponding five spatial RSE and EPSE. For horizon $\phi = 0$ these five RSE and EPSE are contemporaneous, and for $\phi = 1, ..., T-1$ they are dynamic. Below we use one of the total cost and total revenue RSE and EPSE (i.e., $^{\text{Tot}}_{\text{In},i\phi}$) to set out the methods for all five cases. This is because it is simple to modify our presentation to the other four spatial cost and revenue RSE and EPSE by replacing $^{\text{Tot}}_{\text{In},i\phi}$ with the notation for direct ($^{\text{Dir}}_{i\phi}$), indirect spill-in ($^{\text{Ind}}_{\text{In},i\phi}$), etc.

To aid the interpretations of the five spatial revenue (cost) RSE and EPSE and the related two remaining research questions that follow, we assume for simplicity that a bank's input (output) space is two-dimensional and consider the pair of inputs (outputs) m_{it1}, m_{it2} (q_{it1}, q_{it2}). The RSE measures assume that m_{it1}, m_{it2} (q_{it1}, q_{it2}) change equiproportionally along a radial ray, whereas the EPSE measures consider incremental changes in m_{it1}, m_{it2} (q_{it1}, q_{it2}) along a bank's input (output) expansion-path.

(i) $\text{RSE}^{\text{Dir}}_{i\phi}$ and $\text{EPSE}^{\text{Dir}}_{i\phi}$ measure the impact in horizon ϕ of a change in m_{it1}, m_{it2} (q_{it1}, q_{it2}) on the *ith* bank's direct revenue (cost) that is, its own revenue (cost) plus any feedback to this revenue (cost). (ii) $\text{RSE}^{\text{Ind}}_{\text{Out},i\phi}$ and $\text{EPSE}^{\text{Ind}}_{\text{Out},i\phi}$ measure the indirect spill-out impact in ϕ of a change in m_{it1}, m_{it2} (q_{it1}, q_{it2}) on the revenues (costs) of all the other banks in the sample. (iii) $\text{RSE}^{\text{Ind}}_{\text{In},i\phi}$ and $\text{EPSE}^{\text{Ind}}_{\text{In},i\phi}$ measure the impact in ϕ of a change in m_{it1}, m_{it2} (q_{it1}, q_{it2}) on the *ith* bank's revenue (cost), where this impact is due to an indirect spill-in from all the other banks. (iv) $\text{RSE}^{\text{Tot}}_{\text{Out},i\phi}$ and $\text{EPSE}^{\text{Tot}}_{\text{Out},i\phi}$ measure the impact in ϕ of a change in m_{it1}, m_{it2} (q_{it1}, q_{it2}) on the first of the *ith* bank's total revenue (cost) measures, where the two drivers of this impact are the direct impact from (i) and the indirect spill-out impact from (ii). (v) $\text{RSE}^{\text{Tot}}_{\text{In},i\phi}$ and $\text{EPSE}^{\text{Tot}}_{\text{In},i\phi}$ measure the impact in ϕ of a change in m_{it1}, m_{it2} (q_{it1}, q_{it2}) on the second measure of the *ith* bank's total revenue (cost), where the two drivers of this impact are (i) and (iii) above.

Glass, Kenjegaliev, and Kenjegalieva (2020) use a static spatial cost model and therefore analyse only contemporaneous spatial cost returns to scale that is, when $\phi = 0$. In particular, they consider $\text{RSE}^{\text{Dir}}_{i0}$ and $\text{EPSE}^{\text{Dir}}_{i0}$ because they closely resemble standard own returns to scale, and they consider $\text{RSE}^{\text{Ind}}_{\text{In},i0}$ and $\text{EPSE}^{\text{Ind}}_{\text{In},i0}$ because from a bank's own perspective it is the cost spill-ins that matter (and not the cost spill-outs from the bank). To account for both $\text{RSE}^{\text{Dir}}_{i0}$ and $\text{RSE}^{\text{Ind}}_{\text{In},i0}$, and $\text{EPSE}^{\text{Dir}}_{i0}$ and $\text{EPSE}^{\text{Ind}}_{\text{In},i0}$, they also consider $\text{RSE}^{\text{Tot}}_{\text{In},i0}$ and $\text{EPSE}^{\text{Tot}}_{\text{In},i0}$. From a fitted model for large and medium-sized U.S. banks they report, among other things, estimates of the aforementioned spatial returns to scale for the mean large bank (i.e., outside the sample mean). For large U.S. banks they find that $\text{RSE}^{\text{Dir}}_{i0}$ and $\text{EPSE}^{\text{Dir}}_{i0}$ are not significantly different from constant returns, and that $\text{RSE}^{\text{Ind}}_{\text{In},i0}$ and $\text{EPSE}^{\text{Ind}}_{\text{In},i0}$ are not significantly different from zero and constant returns, respectively. Note that the big difference between these $\text{RSE}^{\text{Ind}}_{\text{In},i0}$ and $\text{EPSE}^{\text{Ind}}_{\text{In},i0}$ estimates is because RSE and EPSE measure different things. Based on what we noted above, RSE assume that outputs change equiproportionally along a radial ray, while EPSE allow for the possibility that a bank's output expansion-path is not a radial ray and consider incremental changes along this path. The upshot is that for different reasons they observe that $\text{RSE}^{\text{Tot}}_{\text{In},i0}$ and $\text{EPSE}^{\text{Tot}}_{\text{In},i0}$ are not significantly different from constant returns. As they therefore reach the same policy conclusion using both these total measures – namely, size caps on large banks will, on average, lead to these banks being sub-optimally small and therefore using society's resources inefficiently to provide their services – they do not indicate a preference between spatial RSE and EPSE. For a different data sample, such as the one we use comprising only large U.S. banks, different estimates of corresponding spatial RSE and EPSE may point to different policy recommendations. This is the focus of the following second research question, where the first part of this question also applies to the standard non-spatial RSE and EPSE (RQ2).

Further extensions of Glass, Kenjegaliev, and Kenjegalieva (2020) include looking at optimal bank size through a different lens by computing contemporaneous spatial revenue returns to scale, while also considering the contemporaneous impact of spatial interdependencies from an industry perspective. The latter involves accounting for the cost spill-outs from the banks by computing $\text{RSE}^{\text{Ind}}_{\text{Out},i0}$ and $\text{EPSE}^{\text{Ind}}_{\text{Out},i0}$ and hence $\text{RSE}^{\text{Tot}}_{\text{Out},i0}$ and $\text{EPSE}^{\text{Tot}}_{\text{Out},i0}$. Our principal extension, however, is to account for dynamic spatial interactions. This leads to the following third, and final, research question (RQ3).

The type of situation we have in mind in RQ3 is when a contemporaneous total returns to scale measure (e.g., $\text{EPSE}^{\text{Ind}}_{\text{In},i0}$) points to one returns to scale classification (e.g., increasing returns), while the corresponding dynamic total returns to scale measure for one or more of the next few in-sample periods points to a different classification (e.g., decreasing returns). The contemporaneous measure indicates that it is optimal for the bank to increase its size in the current period, while the dynamic measure indicates that this would move it further away from its optimal size in one or more of the next few in-sample periods. These contemporaneous and dynamic returns to scale measures are therefore inconsistent with one another, so in the empirical analysis we make a recommendation for banks in this type of situation.

Turning to the methods for the spatial revenue and cost $\text{RSE}^{\text{Tot}}_{\text{In},i\phi}$ and $\text{EPSE}^{\text{Tot}}_{\text{In},i\phi}$.

4.4.1 | Spatial revenue and cost RSE

We compute $\text{RSE}^{\text{Tot}}_{\text{In},i\phi}(r)$ as follows and $\text{RSE}^{\text{Tot}}_{\text{In},i\phi}(c)$ using Equation 9.

$$\text{RSE}^{\text{Tot}}_{\text{In},i\phi}(r) = \sum_{g=1}^{G} \frac{\partial r^{\text{Tot}}_{\text{In},i\phi}(t, \mathbf{m}, \mathbf{p})_{it}}{\partial m_{git}}, \tag{8}$$

where the inputs are indexed $g \in 1, ..., G$. The elasticity $\partial r^{\text{Tot}}_{\text{In},i\phi}(t, \mathbf{m}, \mathbf{p})_{it}/\partial m_{git}$ is the first order derivative of the translog function for $r^{\text{Tot}}_{\text{In},i\phi}$ (Equation 7) with respect to input g.

$$\text{RSE}^{\text{Tot}}_{\text{In},i\phi}(c) = \sum_{h=1}^{H} \frac{\partial c^{\text{Tot}}_{\text{In},i\phi}(t, \mathbf{q}, \mathbf{s})_{it}}{\partial q_{hit}}, \tag{9}$$

where the outputs are indexed $h \in 1, ..., H$. The elasticity $\partial c^{\text{Tot}}_{\text{In},i\phi}(t, \mathbf{q}, \mathbf{s})_{it}/\partial q_{hit}$ is the first order derivative of the translog function for $c^{\text{Tot}}_{\text{In},i\phi}$ with respect to output h.

According to production theory, own revenue (cost) RSE should be positive and an estimate $<, =$ or >1 indicates decreasing (increasing), constant or increasing (decreasing) returns to scale. However, there is no theory to indicate whether any of $\text{RSE}^{\text{Dir}}_{i\phi}$, $\text{RSE}^{\text{Ind}}_{\text{In},i\phi}$, $\text{RSE}^{\text{Ind}}_{\text{Out},i\phi}$, $\text{RSE}^{\text{Tot}}_{\text{In},i\phi}$, and $\text{RSE}^{\text{Tot}}_{\text{Out},i\phi}$ should be positive or negative. Should any of these five spatial revenue (cost) RSE measures be positive, then the classification of returns to scale is as above for the standard non-spatial case. Should any of these five spatial revenue (cost) RSE measures be negative, then an estimate $<, =$ or > -1 indicates decreasing (increasing), constant or increasing (decreasing) returns to scale. The classification of the five spatial revenue (cost) RSE measures need not of course be the same.

The five spatial revenue (cost) RSE measures are partially or entirely made up of a spill-in/spill-out. Of these five spatial RSE measures, only $\text{RSE}^{\text{Ind}}_{\text{In},i\phi}$ and $\text{RSE}^{\text{Ind}}_{\text{Out},i\phi}$ are entirely made up of a spill-in / spill-out. Therefore, it is the sign and magnitude of this spill-in / spill-out that is of interest. $\text{RSE}^{\text{Dir}}_{i\phi}$ is made up of own RSE and feedback RSE. As we noted above, from production theory own RSE should be positive, but there is no theory to indicate whether feedback RSE should be positive or negative. If the feedback RSE measure is negative, the issue is whether it is sufficiently negative to make the $\text{RSE}^{\text{Dir}}_{i\phi}$ negative. As the feedback parameter estimates in the empirical spatial literature are small (e.g., Autant-Bernard & LeSage, 2011), we would expect this to also be the case for the feedback RSE. So, even if the feedback RSE measure is negative, $\text{RSE}^{\text{Dir}}_{i\phi}$ would likely be relatively large and positive. As a result of the above method to calculate the five spatial RSE, and a total elasticity being the sum of the direct and indirect spill-in / spill-out elasticities, $\text{RSE}^{\text{Tot}}_{\text{In},i\phi} = \text{RSE}^{\text{Dir}}_{i\phi} + \text{RSE}^{\text{Ind}}_{\text{In},i\phi}$ and $\text{RSE}^{\text{Tot}}_{\text{Out},i\phi} = \text{RSE}^{\text{Dir}}_{i\phi} + \text{RSE}^{\text{Ind}}_{\text{Out},i\phi}$. Given $\text{RSE}^{\text{Dir}}_{i\phi}$ will likely be large and positive, if $\text{RSE}^{\text{Ind}}_{\text{In},i\phi}$ and $\text{RSE}^{\text{Ind}}_{\text{Out},i\phi}$ are negative, the issue is whether they are sufficiently negative to make $\text{RSE}^{\text{Tot}}_{\text{In},i\phi}$ and $\text{RSE}^{\text{Tot}}_{\text{Out},i\phi}$ negative.

The economic explanations for the signs of the $\text{RSE}^{\text{Ind}}_{\text{In},i\phi}$ and $\text{RSE}^{\text{Ind}}_{\text{Out},i\phi}$ measures are the same as the explanations we gave for the sign of a SAR parameter in the above discussion of RQ1; namely, a positive sign is due to banks facing common economic phenomena, while a negative sign is attributed to the effects of spatial competition.

The statistical inference involves using the parameters from the Monte Carlo simulations to compute 1, 000 estimates of each spatial RSE measure.

4.4.2 | Spatial revenue and cost EPSE

Own and spatial RSE are convenient measures and own RSE is the most widely reported measure in the literature. However, in practice, RSE measures may not be the most appropriate because it is entirely feasible that a bank is not located along a radial ray. To overcome this issue, Berger et al. (1987) propose a measure of own scale economies (EPSE) along a bank's non-radial input (output) expansion-path.

To adapt the method for the own contemporaneous revenue EPSE to compute the contemporaneous and dynamic spatial revenue EPSE measures, in the $(t, \mathbf{m}, \mathbf{p})_{it}$ space, consider a point $(t^*, \mathbf{m}^*, \mathbf{p}^*)_{it}$ that lies somewhere along a non-radial ray. Likewise, to compute the spatial cost EPSE measures, consider a point $(t^*, \mathbf{q}^*, \mathbf{s}^*)_{it}$. $\text{EPSE}^{\text{Tot}}_{\text{In},i\phi}(r)$ quantifies the change in expected $r^{\text{Tot}}_{\text{In},i\phi}$ as a bank moves along the non-radial ray between the points $(t^*, (1-\kappa)\mathbf{m}^*, \mathbf{p}^*)_{it}$ and $(t^*, (1+\kappa)\mathbf{m}^*, \mathbf{p}^*)_{it}$, where κ is a small pre-specified number, the choice of which we discuss below. Along the same lines, $\text{EPSE}^{\text{Tot}}_{\text{In},i\phi}(c)$ quantifies the change in expected $c^{\text{Tot}}_{\text{In},i\phi}$ as a bank moves along the non-radial ray between the points $(t^*, (1-\kappa)\mathbf{q}^*, \mathbf{s}^*)_{it}$ and $(t^*, (1+\kappa)\mathbf{q}^*, \mathbf{s}^*)_{it}$.

Using the translog total revenue function in Equation (7) and the corresponding total cost function, we compute $\text{EPSE}^{\text{Tot}}_{\text{In},i\phi}(r)$ and $\text{EPSE}^{\text{Tot}}_{\text{In},i\phi}(c)$ as follows.

$$\begin{aligned}\text{EPSE}^{\text{Tot}}_{\text{In},i\phi}(r) &= \frac{r^{\text{Tot}}_{\text{In},i\phi}(t^*, \vartheta(1-\kappa)\mathbf{m}^*, \mathbf{p}^*)_{it}}{\vartheta r^{\text{Tot}}_{\text{In},i\phi}(t^*, (1-\kappa)\mathbf{m}^*, \mathbf{p}^*)_{it}} \\ &= \frac{r^{\text{Tot}}_{\text{In},i\phi}(t^*, (1+\kappa)\mathbf{m}^*, \mathbf{p}^*)_{it}}{\left(\frac{1+\kappa}{1-\kappa}\right) r^{\text{Tot}}_{\text{In},i\phi}(t^*, (1-\kappa)\mathbf{m}^*, \mathbf{p}^*)_{it}},\end{aligned} \tag{10}$$

$$\text{EPSE}^{\text{Tot}}_{\text{In},i\phi}(c) = \frac{c^{\text{Tot}}_{\text{In},i\phi}(t^*, \vartheta(1-\kappa)\mathbf{q}^*, \mathbf{s}^*)_{it}}{\vartheta c^{\text{Tot}}_{\text{In},i\phi}(t^*, (1-\kappa)\mathbf{q}^*, \mathbf{s}^*)_{it}}$$

$$= \frac{c^{\text{Tot}}_{\text{In},i\phi}(t^*, (1+\kappa)\mathbf{q}^*, \mathbf{s}^*)_{it}}{\left(\frac{1+\kappa}{1-\kappa}\right) c^{\text{Tot}}_{\text{In},i\phi}(t^*, (1-\kappa)\mathbf{q}^*, \mathbf{s}^*)_{it}}. \quad (11)$$

Due to the relative proportions of the *ith* bank's G inputs being constant, $\vartheta(1-\kappa)\mathbf{m}^*_{it} = (1+\kappa)\mathbf{m}^*_{it}$. As this is also the case for the *ith* bank's H outputs, $\vartheta(1-\kappa)\mathbf{q}^*_{it} = (1+\kappa)\mathbf{q}^*_{it}$. This gives $\vartheta = (1+\kappa)/(1-\kappa)$ and thus Equations (10) and (11).

We follow Wheelock and Wilson's (2012) non-spatial study of returns to scale in U.S. banking by using $\kappa = 0.05$. We compute the spatial revenue (cost) EPSE measures for movements along the input (output) expansion-path between $\pm\kappa$ of the mean input (output) vector for the full sample or a subsample. In other words, we consider movements between 95% and 105% of the relevant mean input (output) vector.

As we noted above for the spatial RSE measures, there is no production theory that suggests whether any of the five spatial EPSE measures should be positive or negative. Thus, the returns to scale classification of the five spatial revenue (cost) EPSE measures is the same as the above classification for the corresponding spatial RSE. Therefore, for a positive contemporaneous / dynamic spatial revenue (cost) EPSE measure, an estimate $<, =$ or >1 indicates decreasing (increasing), constant or increasing (decreasing) returns to scale along the specified portion of the relevant expansion-path. Should any of the contemporaneous / dynamic spatial revenue (cost) EPSE measures be negative, then an estimate $<, =$ or > -1 indicates decreasing (increasing), constant or increasing (decreasing) returns to scale. As we noted above for the spatial RSE, the classification of the five spatial EPSE measures need not be the same. The two economic explanations that we gave above for positive and negative spatial RSE are also applicable to explain the sign of a spatial EPSE estimate. Moreover, following the above approach for the spatial RSE, statistical inference for the spatial EPSE for horizon ϕ involves computing 1, 000 estimates.

In contrast to the case of the spatial RSE, $\text{EPSE}^{\text{Tot}}_{\text{In},i\phi}$ and $\text{EPSE}^{\text{Tot}}_{\text{Out},i\phi}$ are not the sum of the direct and indirect spill-in / spill-out EPSE. This is because the five spatial EPSE measures are ratios with different denominators. $\text{EPSE}^{\text{Tot}}_{\text{In},i\phi}$ does though incorporate indirect spill-in returns to a bank from all the other banks in the sample; while $\text{EPSE}^{\text{Tot}}_{\text{Out},i\phi}$ incorporates indirect spill-out returns from the bank to all the other banks.

5 | EMPIRICAL RESULTS AND DISCUSSION

5.1 | Data

Glass and Kenjegalieva (2019) estimate a static spatial cost function for large U.S. banks over the period 1992–2015. As we pointed out in Section 3 the issues they explore differ from those we consider. That is, rather that focus on dynamic returns to scale spillovers as we do here, they compute the growth in static returns to scale spillovers as part of their focus on the decomposition of static spatial total factor productivity growth. Their study period was able go back to 1992 as they use distances between pairs of bank headquarters to specify their spatial weights matrix. We, on the other hand, use richer branch location information to specify the spatial weights matrix (see further in this subsection for a discussion of the a priori construction of this matrix). This branch location information goes back to 1994, while the Interstate Banking and Branching Efficiency (IBBE) Act, otherwise known as the Riegle-Neal Act, came into effect on June 1, 1997. This Act allowed a bank to open branches outside its state of origin and presented greater opportunities for banks to have overlapping branch networks, leading to greater potential spatial dependence between banks. Many states implemented this Act in advance of the effective date. See Table 1 in Dick (2006) for the date when each state implemented the Act. For example, Oregon was the first contiguous state to do so on 2/27/95, while 12 states were the last to implement it on 6/1/97.[ix]

For large and medium-sized U.S. banks over the period 1998–2015, and using branch location information to specify the spatial weights matrix, Glass, Kenjegaliev, and Kenjegalieva (2020); Glass, Kenjegalieva, and Douch (2020) estimate static spatial cost and revenue models, respectively. Glass, Kenjegalieva, and Douch (2020) find evidence of a time lag between when states began implementing the IBBE Act and the spatial dependence between banks that we associate with overlapping branch networks. As we also use branch location information to specify the spatial weights matrix, we follow Glass, Kenjegaliev, and Kenjegalieva (2020); Glass, Kenjegalieva, and Douch (2020) and use a study period that begins in 1998, as this allows time for a sufficient overlap of branch networks and thus spatial dependence to materialize. We use an updated study period of 1998–2019, which, as is also the case for the study period in Glass, Kenjegaliev, and Kenjegalieva (2020); Glass, Kenjegalieva, and Douch (2020), is an interesting period as it includes different bank operating environments, such as, among others, the 2008 crisis. On the evolution over time of spatial returns

to scale for U.S. banks (albeit only static returns), see Glass, Kenjegaliev, and Kenjegalieva (2020). To clearly distinguish our paper from theirs we provide a detailed, policy focused empirical analysis that considers the entirely different issue of the implications of size caps. That is, we apply the methods we introduce for dynamic internal and external returns to scale to the costs and revenues of key large U.S. banks; namely, G-SIBs and CCAR banks.

By applying Berger and Roman's (2017) large bank size total assets threshold to 2015, we classify a U.S. bank as large if its total assets are greater than $3 billion. This yields data comprising 201 large U.S. banks, which because the panel data is balanced represent the core group of surviving large banks. In 2019, the total assets of these banks accounted for 73.5% of the total assets in the U.S. banking system and these banks had 52, 336 branches. In terms of the charter status of the banks, the sample comprises 55 Fed member federal charter commercial banks; 43 state charter commercial or savings Fed member banks; 75 state charter Fed nonmember commercial banks; 16 savings banks; and 12 savings associations. 86.1% of the banks in the sample (173 banks) specialize in commercial lending. As we noted in Section 2, we focus on large banks because their branch networks are sufficiently large and so there is a more than sufficient overlap between their networks that is, there is not a lack of interconnectedness between the banks in our spatial weights matrix.

TABLE 1 Variable descriptions and descriptive statistics

Variable description	Model notation	Mean	SD
Dependent variables			
Total operating cost (000 s of 2005 U.S. dollars): Sum of salaries, interest expenses on deposits and expenditure on fixed assets and premises	c	779, 698	3, 319, 010
Total revenue measure (000 s of 2005 U.S. dollars): Interest income plus non-interest income	r	1, 844, 788	7, 984, 531
Input prices and outputs in the cost model			
Cost of fixed assets and premises: Expenditure on fixed assets and premises divided by their value	s_1	1.05	20.07
Cost of labor: Salaries divided by the number of full-time equivalent employees	s_2	66.09	22.79
Cost of deposits: Interest expenses on deposits divided by total deposits	s_3	0.02	0.01
Net loans and leases (000 s of 2005 U.S. dollars)	q_1	17, 543, 583	73, 260, 471
Securities (000 s of 2005 U.S. dollars)	q_2	6, 449, 827	29, 147, 718
Non-interest income (000s of 2005 U.S. dollars)	q_3	615, 346	2, 886, 586
Output prices and inputs in the revenue model			
Price of loans and leases: Interest income from loans and leases divided by loans and leases	p_1	0.057	0.016
Price of securities: Interest income from securities divided by securities	p_2	0.038	0.028
Price of other activities: Approximated by non-interest income divided by total assets	p_3	0.013	0.016
Fixed assets and premises: Value in 000s of 2005 U.S. dollars	m_1	269, 847	996, 639
Labor: number of full-time equivalent employees	m_2	5, 664	22, 953
Deposits in 000s of 2005 U.S. dollars	m_3	23, 956, 077	108, 700, 007

The data for the variables is from the Call Reports and was sourced from the Federal Deposit Insurance Corporation (FDIC). This data is at the bank level and having based the general categorization of a variable as an input or an output on the well-established intermediation approach to banking (Sealey & Lindley, 1977) – e.g., this approach regards deposits as an input—the specific measures of the outputs and inputs (as well as their prices) are based on those in Koetter et al. (2012).[x] See Table 1 for a description of the measures of the first order outputs and inputs (as well as their prices), along with the summary statistics. It, of course, follows from the translog functional form that we also include squared and interaction terms pertaining to the time trend, t, outputs and inputs (and their prices).

In summary, there are three input prices and three outputs in the cost model and hence, three output prices and three inputs in the revenue model. The outputs (and their prices) relate to net loans and leases, q_1 (p_1), securities, q_2 (p_2) and non-interest income, q_3 (p_3). The inputs (and their prices) relate to fixed assets and premises, m_1 (s_1), labor, m_2 (s_2) and deposits, m_3 (s_3). c denotes total operating cost and is the sum of the expenditures on the inputs, and r denotes the total revenue measure – interest income plus non-interest income. We deflate c, r, $q_1 - q_3$, m_1 and m_3 to 2005 prices using the CPI, but not the input and output prices because, as we can see from Table 1, they are ratios. All the variables are then logged, mean adjusted and, finally, we use s_1 as the normalizing factor for c and the other input prices, and p_1 as the normalizing factor for r and the other output prices. By mean adjusting the data, the contemporaneous and dynamic direct, indirect and total parameters on the first order variables are elasticities at the sample mean. From Table 1 we can see for a number of variables that the standard deviation is rather large, vis-à-vis the mean. Such dispersion between the observations of a variable is because there are a relatively small number of very large banks in the sample.

Using the state locations of each bank's branches in the Summary of Deposits from the FDIC, we specify the same $\mathbf{W}$ for the cost and revenue models in four steps.

1. Set about obtaining a $\mathbf{W}$ before normalization for each year by setting all the cells on the main diagonals of these annual matrices to zero. This is because a bank cannot be its own neighbour.
2. For each state where the ith and jth banks have branches in year t, we calculate the ratio of the number of jth bank branches to the number of ith bank branches. We then sum these ratios across the states where the ith and jth banks have branches to obtain the non-zero off-diagonal elements of the annual matrices. All the other off-diagonal elements in the annual matrices are set to zero to signify that the corresponding ith and jth banks' branch networks do not overlap. Relative to the branch network of the ith bank, each of the off-diagonal elements can be interpreted as the relative branch network intensity of the jth bank.
3. We then average the annual matrices from (2).[xi]
4. We obtain the $\mathbf{W}$ we use in the estimation of the models by normalizing the average matrix from (3) by dividing throughout by its largest cell (i.e., eigenvalue). The advantage of this normalization is that it retains the information on the relative intensities of the banks' branch networks, as it does not change the proportional relationship between the spatial weights.[xii]

5.2 | Estimated dynamic spatial cost and revenue models

In Tables 2 and 3 we present the estimated dynamic spatial cost and revenue models (Equations 5 and 6). Further in this subsection we discuss the estimates of the cumulative indirect parameters which measure the global spillovers pertaining to a bank's 1st order and higher order neighbours. In Tables 2 and 3, the estimates of the coefficients on $\mathbf{W}c_t$ and $\mathbf{W}r_t$ ($\mathbf{W}c_{t-1}$ and $\mathbf{W}r_{t-1}$) are local impacts that measure the contemporaneous (dynamic) cost and revenue spillovers to a bank from only its 1st order neighbours.

From Table 2 we can see that the estimates of the coefficients on $\mathbf{W}c_t$ and $\mathbf{W}c_{t-1}$ are positive, non-negligible and significant at the 5% level or lower. These positive contemporaneous and dynamic local cost spillovers are consistent with neighbouring 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. In particular, we observe a non-negligible difference between the larger $\mathbf{W}c_{t-1}$ parameter and the coefficient on $\mathbf{W}c_t$. This points to it taking some time for the larger of these two local cost spillovers to occur. Our finding that the $\mathbf{W}c_t$ parameter is positive, non-negligible and significant is very much in line with the extant evidence for U.S. banks. This evidence is from Glass and Kenjegalieva (2019) for large banks and Glass, Kenjegaliev, and Kenjegalieva (2020) for large and medium-sized banks. However, our paper is the first to apply a dynamic spatial cost model to U.S. banks and finds that the $\mathbf{W}c_{t-1}$ variable (which by construction was omitted in the above authors' static spatial models) is a particularly important regressor for our application.

TABLE 2 Estimated dynamic spatial cost model

	Model coeff.		Model coeff.		Model coeff.
$\mathbf{W}c_t$	0.205***	t^2	0.000	$\mathbf{W}q_{2t}s_{3t}$	0.011
$\mathbf{W}c_{t-1}$	0.260**	$q_{1t}t$	−0.004***	$\mathbf{W}q_{3t}s_{2t}$	0.127
q_{1t}	0.549***	$q_{2t}t$	−0.001	$\mathbf{W}q_{3t}s_{3t}$	−0.011
q_{2t}	0.188***	$q_{3t}t$	0.004***	$\mathbf{W}q_{1t-1}$	−0.259*
q_{3t}	0.179***	$s_{2t}t$	−0.001	$\mathbf{W}q_{2t-1}$	0.148
s_{2t}	0.620***	$s_{3t}t$	−0.003***	$\mathbf{W}q_{3t-1}$	−0.168*
s_{3t}	0.294***	$\mathbf{W}q_{1t}$	0.229*	$\mathbf{W}s_{2t-1}$	−0.310
q_{1t}^2	0.062***	$\mathbf{W}q_{2t}$	−0.134	$\mathbf{W}s_{3t-1}$	−0.056
q_{2t}^2	0.023***	$\mathbf{W}q_{3t}$	−0.228**	$\mathbf{W}q_{1t-1}^2$	0.286***
q_{3t}^2	0.044***	$\mathbf{W}s_{2t}$	−0.350**	$\mathbf{W}q_{2t-1}^2$	0.015
$q_{1t}q_{2t}$	−0.050***	$\mathbf{W}s_{3t}$	−0.066	$\mathbf{W}q_{3t-1}^2$	0.058
$q_{1t}q_{3t}$	−0.073***	$\mathbf{W}q_{1t}^2$	−0.195**	$\mathbf{W}q_{1t-1}q_{2t-1}$	−0.267***
$q_{2t}q_{3t}$	−0.003	$\mathbf{W}q_{2t}^2$	−0.036	$\mathbf{W}q_{1t-1}q_{3t-1}$	−0.247**
s_{2t}^2	0.041***	$\mathbf{W}q_{3t}^2$	0.012	$\mathbf{W}q_{2t-1}q_{3t-1}$	0.163**
s_{3t}^2	0.036***	$\mathbf{W}q_{1t}q_{2t}$	0.182*	$\mathbf{W}s_{2t-1}^2$	−0.101
$s_{2t}s_{3t}$	−0.077***	$\mathbf{W}q_{1t}q_{3t}$	0.099	$\mathbf{W}s_{3t-1}^2$	−0.010
$q_{1t}s_{2t}$	0.066***	$\mathbf{W}q_{2t}q_{3t}$	−0.073	$\mathbf{W}s_{2t-1}s_{3t-1}$	−0.081
$q_{1t}s_{3t}$	0.033***	$\mathbf{W}s_{2t}^2$	−0.120	$\mathbf{W}q_{1t-1}s_{2t-1}$	− 0.214
$q_{2t}s_{2t}$	−0.017**	$\mathbf{W}s_{3t}^2$	−0.001	$\mathbf{W}q_{1t-1}s_{3t-1}$	−0.051
$q_{2t}s_{3t}$	0.021***	$\mathbf{W}s_{2t}s_{3t}$	0.090	$\mathbf{W}q_{2t-1}s_{2t-1}$	−0.046
$q_{3t}s_{2t}$	−0.016**	$\mathbf{W}q_{1t}s_{2t}$	0.093	$\mathbf{W}q_{2t-1}s_{3t-1}$	0.022
$q_{3t}s_{3t}$	−0.044***	$\mathbf{W}q_{1t}s_{3t}$	0.020	$\mathbf{W}q_{3t-1}s_{2t-1}$	0.220**
t	−0.004***	$\mathbf{W}q_{2t}s_{2t}$	−0.163	$\mathbf{W}q_{3t-1}s_{3t-1}$	0.012

Note: $LL = 3959.6$; *, ** and *** denote statistical significance at the 10%, 5% and 1% levels, respectively.

The estimate of the $\mathbf{W}r_t$ parameter in Table 3 is positive, of moderate magnitude and significant at the 10% level. Finding that this parameter is positive is in line with the extant evidence for large and medium-sized U.S. banks from a static spatial revenue model (Glass, Kenjegalieva, & Douch, 2020) and is consistent with neighbouring banks' revenues being impacted by common contemporaneous economic phenomena. Our paper is the first to apply a dynamic spatial revenue model to U.S. banks and, whereas it is the $\mathbf{W}r_t$ parameters in the Glass, Kenjegalieva, and Douch (2020) model that are non-negligible and significant at the 5% level or lower, we find that it is the $\mathbf{W}r_{t-1}$ variable that has a large, significant (at the 1% level) and, interestingly, negative impact, which provides support for the inclusion of this dynamic SAR variable.[xiii] This negative impact is by far the largest of the two SAR revenue dependencies we consider and is consistent with it taking some time for the diffusion of competitive effects across space to take effect. Additionally, it is evident from Tables 2 and 3 that some of the $\mathbf{WX}_t$ and $\mathbf{WX}_{t-1}$ variables are significant at the 10% level or lower. These results are supportive of our dynamic SDM specification, as opposed to a dynamic SAR model which would omit $\mathbf{WX}_t$ and $\mathbf{WX}_{t-1}$.

In Tables 2 and 3 the coefficients on the first order outputs ($q_{1t} - q_{3t}$), input prices (s_{2t} and s_{3t}), inputs ($m_{1t} - m_{3t}$) and output prices (p_{2t} and p_{3t}) are own elasticities at the sample mean. To satisfy at the sample mean the monotonicity properties of the functions from production theory, these elasticities must be positive, which is the case for the output and input price elasticities from the fitted cost function.

However, although the m_{2t}, m_{3t}, p_{2t} and p_{3t} parameters from the fitted revenue function are positive, the m_{1t} (fixed assets and premises) parameter is negative and significant at the 1% level, although its magnitude is extremely small. Despite the negative sign of this parameter being at odds with the theoretical monotonicity property that underpins a revenue function, fixed assets and premises is retained as an input in the model. This is because there is some evidence to suggest that for a large portion of our sample the own m_{1t} elasticity at the sample

TABLE 3 Estimated dynamic spatial revenue model

	Model coeff.		Model coeff.		Model coeff.
$\mathbf{W}r_t$	0.077*	t^2	0.0002***	$\mathbf{W}m_{2t}p_{3t}$	0.137
$\mathbf{W}r_{t-1}$	−0.364***	$m_{1t}t$	0.003***	$\mathbf{W}m_{3t}p_{2t}$	0.138
m_{1t}	−0.023***	$m_{2t}t$	−0.006***	$\mathbf{W}m_{3t}p_{3t}$	0.000
m_{2t}	0.113***	$m_{3t}t$	0.002*	$\mathbf{W}m_{1t-1}$	0.252**
m_{3t}	0.902***	$p_{2t}t$	−0.009***	$\mathbf{W}m_{2t-1}$	−0.062
p_{2t}	0.143***	$p_{3t}t$	−0.002***	$\mathbf{W}m_{3t-1}$	0.244
p_{3t}	0.193***	$\mathbf{W}m_{1t}$	0.050	$\mathbf{W}p_{2t-1}$	0.194***
m_{1t}^2	−0.011***	$\mathbf{W}m_{2t}$	−0.370**	$\mathbf{W}p_{3t-1}$	0.059
m_{2t}^2	0.009	$\mathbf{W}m_{3t}$	0.246	$\mathbf{W}m_{1t-1}^2$	−0.043
m_{3t}^2	0.000	$\mathbf{W}p_{2t}$	0.268***	$\mathbf{W}m_{2t-1}^2$	−0.190
$m_{1t}m_{2t}$	0.024**	$\mathbf{W}p_{3t}$	0.042	$\mathbf{W}m_{3t-1}^2$	−0.154
$m_{1t}m_{3t}$	−0.004	$\mathbf{W}m_{1t}^2$	−0.063	$\mathbf{W}m_{1t-1}m_{2t-1}$	0.071
$m_{2t}m_{3t}$	−0.018	$\mathbf{W}m_{2t}^2$	0.218	$\mathbf{W}m_{1t-1}m_{3t-1}$	−0.004
p_{2t}^2	−0.018***	$\mathbf{W}m_{3t}^2$	0.075	$\mathbf{W}m_{2t-1}m_{3t-1}$	0.313
p_{3t}^2	0.053***	$\mathbf{W}m_{1t}m_{2t}$	−0.090	$\mathbf{W}p_{2t-1}^2$	0.051
$p_{2t}p_{3t}$	−0.008	$\mathbf{W}m_{1t}m_{3t}$	0.194	$\mathbf{W}p_{3t-1}^2$	−0.001
$m_{1t}p_{2t}$	0.000	$\mathbf{W}m_{2t}m_{3t}$	−0.349*	$\mathbf{W}p_{2t-1}p_{3t-1}$	−0.113*
$m_{1t}p_{3t}$	−0.009**	$\mathbf{W}p_{2t}^2$	−0.123***	$\mathbf{W}m_{1t-1}p_{2t-1}$	−0.151*
$m_{2t}p_{2t}$	0.002	$\mathbf{W}p_{3t}^2$	−0.055	$\mathbf{W}m_{1t-1}p_{3t-1}$	−0.013
$m_{2t}p_{3t}$	0.033***	$\mathbf{W}p_{2t}p_{3t}$	0.096	$\mathbf{W}m_{2t-1}p_{2t-1}$	0.150
$m_{3t}p_{2t}$	−0.008	$\mathbf{W}m_{1t}p_{2t}$	0.016	$\mathbf{W}m_{2t-1}p_{3t-1}$	0.171
$m_{3t}p_{3t}$	−0.020***	$\mathbf{W}m_{1t}p_{3t}$	−0.163*	$\mathbf{W}m_{3t-1}p_{2t-1}$	−0.022
t	0.000	$\mathbf{W}m_{2t}p_{2t}$	−0.191	$\mathbf{W}m_{3t-1}p_{3t-1}$	−0.118

Note: $LL = 5775.5$; *, ** and *** denote statistical significance at the 10%, 5% and 1% levels, respectively.

mean is likely to be positive. This is consistent with the estimate of this elasticity that Glass, Kenjegalieva, and Douch (2020) report from a static spatial revenue model for a larger sample of U.S. banks (both medium-sized and large institutions) over the shorter period 1998–2015. Given this finding, using data for the same 201 banks that we use to estimate the spatial revenue model, we estimate standard non-spatial translog revenue models with fixed effects for 1998–2019 and a series of shorter periods which successively drop the final year. From these non-spatial models we find that the m_{1t} elasticity at the sample mean goes from being small, negative and significant for the samples that end in 2016–2019; to small, negative or positive, and not significant for the samples that end in 2011–2015; and then small, positive and significant for the sample that ends in 2010.[xiv] This suggests that in recent years there has been a change in the nature of the relationship between banks' fixed assets and premises and their revenues, which is consistent with the growth in online banking leading to branch closures.

From the cost and revenue models we also obtain direct, symmetric indirect and symmetric total parameters. Unlike the own parameters, direct, indirect and total parameters are partially or entirely made up of spillovers. In Table 4, from the cost function and for the full set of in-sample time horizons, we present the cumulative direct, indirect and total output and input price elasticities at the sample mean. In Table 5, from the revenue function, we present the corresponding contemporaneous and dynamic elasticities for the inputs and output prices.

We can see from Tables 4 and 5 that the cumulative direct elasticities at the sample mean for all the in-sample time horizons are, with the exception of one estimate, significant at the 5% level or lower.[xv] These direct elasticities are also essentially of the same magnitude as the corresponding own elasticity in Tables 2 and 3 which points to two things. First, there are negligible feedback effects within the direct parameters and so in the next subsection the direct returns to scale are akin to own returns. Second, the direct effects in the contemporaneous horizon are very persistent, as the magnitudes of these direct

TABLE 4 Cumulative direct, indirect and total elasticities at the sample mean from the cost function

Horizon, ϕ	q_1 (loans and leases)			q_2 (securities)			q_3 (non-interest income)			s_2 (cost of labor)			s_3 (cost of deposits)		
	Direct	Indirect	Total	Direct	Indirect	Total	Direct	Indirect	Total	Direct	Indirect	Total	Direct	Indirect	Total
0	0.550***	0.098***	0.647***	0.188***	0.062***	0.250***	0.178***	−0.423***	−0.245***	0.619***	−0.628***	−0.009	0.294***	−0.071***	0.222***
1	0.550***	0.312***	0.863***	0.188***	0.150***	0.338***	0.176***	−0.581***	−0.404***	0.617***	−0.755***	−0.139**	0.294***	−0.019	0.275***
2	0.552***	0.444**	0.995***	0.189***	0.204**	0.392***	0.176***	−0.676***	−0.501***	0.616***	−0.832***	−0.215*	0.294***	0.014	0.308***
3	0.552***	0.533*	1.085***	0.189***	0.240*	0.429***	0.175***	−0.741***	−0.566**	0.616***	−0.883***	−0.267	0.294***	0.036	0.330***
4	0.553***	0.599	1.152**	0.189***	0.267	0.456***	0.175***	−0.789**	−0.614**	0.616***	−0.921***	−0.306	0.294***	0.052	0.347***
5	0.553***	0.652	1.205*	0.189***	0.289	0.478**	0.174***	−0.827**	−0.653*	0.615***	−0.952***	−0.336	0.295***	0.065	0.360***
6	0.554***	0.697	1.250	0.189***	0.307	0.497*	0.174***	−0.860*	−0.686	0.615***	−0.978**	−0.363	0.295***	0.077	0.371**
7	0.554***	0.737	1.291	0.190***	0.324	0.513	0.174***	−0.889	−0.715	0.615***	−1.001**	−0.386	0.295***	0.087	0.381*
8	0.554***	0.775	1.330	0.190***	0.339	0.529	0.174***	−0.917	−0.743	0.615***	−1.023*	−0.408	0.295***	0.096	0.391*
9	0.554***	0.813	1.368	0.190***	0.355	0.545	0.173***	−0.944	−0.771	0.615***	−1.045	−0.430	0.295***	0.105	0.400
10	0.555***	0.852	1.407	0.190***	0.371	0.560	0.173***	−0.972	−0.799	0.614***	−1.067	−0.453	0.295***	0.115	0.410
11	0.555***	0.893	1.448	0.190***	0.387	0.577	0.173***	−1.001	−0.828	0.614***	−1.090	−0.476	0.295***	0.125	0.420
12	0.555***	0.937	1.492	0.190***	0.405	0.595	0.173***	−1.033	−0.861	0.614***	−1.116	−0.502	0.295***	0.136	0.431
13	0.556***	0.987	1.542	0.190***	0.425	0.616	0.172***	−1.069	−0.897	0.614***	−1.144	−0.530	0.295***	0.148	0.443
14	0.556***	1.043	1.599	0.191***	0.448	0.639	0.172***	−1.110	−0.937	0.613***	−1.176	−0.563	0.295***	0.162	0.457
15	0.557***	1.109	1.665	0.191***	0.475	0.666	0.172***	−1.157	−0.985	0.613***	−1.213	−0.600	0.295***	0.178	0.473
16	0.557***	1.186	1.743	0.191***	0.506	0.697	0.171***	−1.212	−1.041	0.613***	−1.257	−0.644	0.296***	0.196	0.492
17	0.558***	1.279	1.837	0.191***	0.544	0.735	0.171***	−1.278	−1.107	0.612***	−1.309	−0.697	0.296***	0.219	0.514
18	0.559***	1.391	1.950	0.192***	0.589	0.781	0.170***	−1.358	−1.188	0.612***	−1.372	−0.760	0.296***	0.246	0.542
19	0.560***	1.528	2.088	0.192***	0.644	0.837	0.169***	−1.455	−1.286	0.611***	−1.449	−0.838	0.296***	0.279	0.575
20	0.562***	1.697	2.259	0.193***	0.713	0.905	0.168**	−1.575	−1.407	0.611***	−1.544	−0.933	0.297***	0.319	0.616
21	0.563***	1.907	2.471	0.193***	0.797	0.991	0.167	−1.724	−1.557	0.610***	−1.661	−1.051	0.297***	0.370	0.667

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

TABLE 5 Cumulative direct, indirect and total elasticities at the sample mean from the revenue function

Horizon, ϕ	m_1 (fixed assets and premises)			m_2 (labor)			m_3 (deposits)			p_2 (price of securities)			p_3 (price of other activities)		
	Direct	Indirect	Total	Direct	Indirect	Total	Direct	Indirect	Total	Direct	Indirect	Total	Direct	Indirect	Total
0	−0.023***	0.282***	0.259***	0.112***	−0.399***	−0.287***	0.903***	0.528***	1.431***	0.144***	0.445***	0.589***	0.193***	0.109***	0.302***
1	−0.024***	0.129***	0.105**	0.114***	−0.210***	−0.095	0.900***	−0.081	0.819***	0.141***	0.142	0.284***	0.192***	−0.019	0.173***
2	−0.023***	0.224***	0.200***	0.113***	−0.327***	−0.213***	0.903***	0.299***	1.203***	0.143***	0.331***	0.473***	0.193***	0.061***	0.254***
3	−0.024***	0.161***	0.137***	0.114***	−0.249***	−0.135**	0.901***	0.046	0.947***	0.142***	0.205**	0.347***	0.192***	0.008	0.200***
4	−0.023***	0.205***	0.182***	0.113***	−0.304***	−0.191***	0.903***	0.225**	1.128***	0.142***	0.294***	0.437***	0.193***	0.046**	0.238***
5	−0.024***	0.172***	0.148***	0.114***	−0.263***	−0.149***	0.902***	0.091	0.992***	0.142***	0.228**	0.369***	0.193***	0.017	0.210***
6	−0.024***	0.198***	0.175***	0.113***	−0.295***	−0.182***	0.903***	0.197	1.099***	0.142***	0.280***	0.422***	0.193***	0.040	0.232***
7	−0.024***	0.176***	0.153***	0.114***	−0.268***	−0.155***	0.902***	0.109	1.011***	0.142***	0.237**	0.379***	0.193***	0.021	0.214***
8	−0.024***	0.195***	0.172***	0.113***	−0.291***	−0.178***	0.903***	0.185	1.087***	0.142***	0.274***	0.416***	0.193***	0.037	0.230***
9	−0.024***	0.178***	0.155***	0.114***	−0.271***	−0.157**	0.902***	0.117	1.019***	0.142***	0.241**	0.383***	0.193***	0.023	0.215***
10	−0.024***	0.194***	0.171***	0.113***	−0.290***	−0.177**	0.902***	0.180	1.083***	0.142***	0.272**	0.414***	0.193***	0.036	0.229***
11	−0.024***	0.179**	0.155**	0.114***	−0.271***	−0.158*	0.902***	0.119	1.021***	0.142***	0.242*	0.384***	0.193***	0.023	0.216***
12	−0.024***	0.194**	0.171**	0.113***	−0.290***	−0.177*	0.902***	0.180	1.083***	0.142***	0.272*	0.414***	0.193***	0.036	0.229***
13	−0.024***	0.179*	0.155	0.114***	−0.271**	−0.157	0.902***	0.117	1.019**	0.142***	0.241	0.383*	0.193***	0.023	0.215**
14	−0.024***	0.195	0.171	0.113***	−0.291*	−0.178	0.903***	0.184	1.086**	0.142***	0.274	0.416*	0.193***	0.037	0.230**
15	−0.024***	0.177	0.153	0.114***	−0.269	−0.155	0.902***	0.112	1.014	0.142***	0.238	0.380	0.193***	0.022	0.214
16	−0.024***	0.197	0.173	0.113***	−0.293	−0.180	0.903***	0.191	1.094	0.142***	0.277	0.420	0.193***	0.039	0.231
17	−0.024***	0.175	0.151	0.114***	−0.266	−0.152	0.902***	0.101	1.003	0.142***	0.233	0.375	0.193***	0.019	0.212
18	−0.024***	0.200	0.177	0.113***	−0.297	−0.184	0.903***	0.205	1.107	0.142***	0.284	0.426	0.193***	0.041	0.234
19	−0.024***	0.171	0.147	0.114***	−0.261	−0.147	0.902***	0.085	0.987	0.142***	0.225	0.367	0.192***	0.016	0.208
20	−0.024***	0.205	0.182	0.113***	−0.304	−0.190	0.903***	0.225	1.128	0.142***	0.294	0.436	0.193***	0.046	0.239
21	−0.024***	0.164	0.140	0.114***	−0.253	−0.139	0.901***	0.059	0.960	0.142***	0.212	0.354	0.192***	0.010	0.203

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

effects essentially remain unchanged in all the future in-sample horizons and, apart from the one aforementioned exception, these effects are significant.

It is evident for each of the variables in Tables 4 and 5 that we observe a significant cumulative indirect elasticity at the sample mean for at least one of the in-sample time horizons. For the cost function variables in Table 4, the significant cumulative indirect elasticities are for horizons ranging from 0 (all variables at the 1% level) to 8 (cost of labor, s_2, at the 10% level). The only significant cumulative indirect elasticity for s_3 (cost of deposits) is for horizon 0, whereas for all the other cost function variables in Table 4 the cumulative indirect elasticities are significant up to at least horizon 3.

For the revenue function variables in Table 5, the significant cumulative indirect elasticities are for horizons ranging from 0 (all variables at the 1% level) to 14 (labor, m_2, at the 5% level). For all the variables in this table we observe significant cumulative indirect elasticities for three or more in-sample time horizons. For the sample average bank this indicates that for each of these variables there is significant contemporaneous and future persistent spillover impacts. Interestingly, the magnitudes of the reported cumulative indirect elasticities for the revenue function variables fall and rise from 1 year to the next. This variability is particularly marked for m_3 (deposits). However, Table 5 shows that such fluctuations die out over time. The reason for such variability is the alternating annual impact on the cumulative indirect elasticities of the positive and negative contemporaneous and dynamic SAR parameters (in particular, the alternating annual impact of these parameters on $\mathbf{L}_\phi$ in Equation 4). In contrast, since the contemporaneous and dynamic SAR parameters are positive in the fitted cost function, the significant cumulative indirect elasticities in Table 4 increase annually.

Summing the corresponding cumulative direct and indirect elasticities in Tables 4 and 5 yields the cumulative total elasticity. From these tables, when the signs of the cumulative direct and indirect elasticities are generally the same (different), we find that the cumulative total elasticities are significant over more (fewer) time horizons than the corresponding indirect elasticity that is, a cumulative total impact is more (less) persistent over time than its indirect counterpart.

5.3 | Contemporaneous and dynamic spatial returns to scale estimates

In panel A of Table 6, from the cost function, we present for the sample average bank and time horizons 0, ..., 5, the cumulative estimates of $\text{RSE}^{\text{Dir}}_{i\phi}(c)$, $\text{RSE}^{\text{Ind}}_{\text{In},i\phi}(c) = \text{RSE}^{\text{Ind}}_{\text{Out},i\phi}(c)$, and $RSE^{Tot}_{In,i\phi}(c) = \text{RSE}^{\text{Tot}}_{\text{Out},i\phi}(c)$. In panels B, C and D of the same table we present the corresponding EPSE (c), RSE (r) and EPSE (r) estimates. Recall that because we observed that there is negligible feedback within the contemporaneous and dynamic direct parameters, the contemporaneous and dynamic direct RSE and EPSE are akin to the standard internal returns to scale that are widely reported in the literature. Hence, when there is an increase in the sample average bank's outputs (inputs) in the current period, the $\text{RSE}^{\text{Dir}}_{i\phi}$ and $\text{EPSE}^{\text{Dir}}_{i\phi}$ in Table 6 measure the percentage changes in the contemporaneous and dynamic cost (revenue) of this hypothetical bank.

For the sample average bank for horizons 0, ..., 5, we can see from Table 6 that the estimates of $\text{RSE}^{\text{Dir}}_{i\phi}(c)$ are all significantly less than 1, whereas with the exception of horizon 4 we cannot reject $\text{EPSE}^{\text{Dir}}_{i\phi}(c) = 1$. In the main therefore, these tests indicate that the $\text{EPSE}^{\text{Dir}}_{i\phi}(c)$ estimates point to constant direct cost returns, which is in contrast to the corresponding increasing $\text{RSE}^{\text{Dir}}_{i\phi}(c)$. The corresponding estimates of $\text{RSE}^{\text{Dir}}_{i\phi}(r)$ and $\text{EPSE}^{\text{Dir}}_{i\phi}(r)$ are all of a very similar magnitude (0.99), which, in practice, points to approximately constant direct revenue returns. Interestingly though, these $\text{RSE}^{\text{Dir}}_{i\phi}(r)$ are all significantly less than 1, while there is also some evidence of this for $\text{EPSE}^{\text{Dir}}_{i\phi}(r)$. Such results are due to these returns having small standard deviations. With reference to the results in the literature, our findings of constant and increasing contemporaneous internal returns to scale are in line with results reported in some of the studies reviewed in Section 3 (e.g., Wheelock & Wilson, 2012, 2018).

From the above analysis of the direct cost and revenue RSE and EPSE we reach two conclusions. First, we conclude that these direct returns are very persistent as their magnitudes in future time horizons are the same as in horizon 0. Second, given an EPSE measure can be viewed as more appropriate than the corresponding RSE as a bank may not lie on a radial ray, and when we consider banks in isolation (i.e., we overlook for the moment returns to scale spillovers between banking networks), the $\text{EPSE}^{\text{Dir}}_{i\phi}(c)$ and $\text{EPSE}^{\text{Dir}}_{i\phi}(r)$ suggest that the size of the sample average bank is approximately optimal in horizons 0–5. The issue is therefore whether we reach the same conclusion from EPSE that account for such spillovers, which is what we now consider.

When there is an increase in the sample average bank's outputs (inputs) in the current period, the $\text{RSE}^{\text{Ind}}_{i\phi}$ and $\text{EPSE}^{\text{Ind}}_{i\phi}$ in Table 6 measure the percentage changes in the contemporaneous and dynamic symmetric cost (revenue) spill-in and spill-out to and from this hypothetical bank. We observe that the $\text{RSE}^{\text{Ind}}_{i\phi}(c)$ and $\text{RSE}^{\text{Ind}}_{i\phi}(r)$ in Table 6 do not consistently have the same sign for horizons 0–5. In contrast, from the same table, we can

TABLE 6 Sample average cumulative spatial returns to scale

Panel A: Cost spatial cumulative ray-scale economies			
Horizon, ϕ	$RSE^{Dir}_{i\phi}(\boldsymbol{c})$	$RSE^{Ind}_{In,i\phi}(\boldsymbol{c}) = RSE^{Ind}_{Out,i\phi}(\boldsymbol{c})$	$RSE^{Tot}_{In,i\phi}(\boldsymbol{c}) = RSE^{Tot}_{Out,i\phi}(\boldsymbol{c})$
0	0.915^{*}_{a}	-0.263^{*}_{a}	0.653^{*}_{a}
1	0.915^{*}_{a}	-0.118^{*}_{a}	0.796^{*}_{a}
2	0.916^{*}_{a}	-0.029^{*}_{a}	0.887^{*}_{a}
3	0.916^{*}_{a}	0.032^{*}_{a}	0.948^{*}_{a}
4	0.917^{*}_{a}	0.077^{*}_{a}	0.994^{*}
5	0.917^{*}_{a}	0.113^{*}_{a}	1.030^{*}
Panel B: Cost spatial cumulative expansion-path scale economies			
Horizon, ϕ	$EPSE^{Dir}_{i\phi}(\boldsymbol{c})$	$EPSE^{Ind}_{In,i\phi}(\boldsymbol{c}) = EPSE^{Ind}_{Out,i\phi}(\boldsymbol{c})$	$EPSE^{Tot}_{In,if}(\boldsymbol{c}) = EPSE^{Tot}_{Out,i\phi}(\boldsymbol{c})$
0	1.010^{*}	0.971^{*}_{a}	1.026^{*}_{b}
1	0.982^{*}	0.988^{*}	0.997^{*}
2	0.990^{*}	0.985^{*}	0.998^{*}
3	1.001^{*}	0.979^{*}	1.002^{*}
4	0.983^{*}_{a}	0.974^{*}	1.012^{*}
5	1.000^{*}	0.964^{*}_{a}	0.992^{*}
Panel C: Revenue spatial cumulative ray-scale economies			
Horizon, ϕ	$RSE^{Dir}_{i\phi}(\boldsymbol{r})$	$RSE^{Ind}_{In,i\phi}(\boldsymbol{r}) = RSE^{Ind}_{Out,i\phi}(\boldsymbol{r})$	$RSE^{Tot}_{In,i\phi}(\boldsymbol{r}) = RSE^{Tot}_{Out,i\phi}(\boldsymbol{r})$
0	0.993^{*}_{a}	0.411^{*}_{a}	1.404^{*}_{b}
1	0.990^{*}_{a}	-0.162^{*}_{a}	0.828^{*}_{a}
2	0.993^{*}_{a}	0.196^{*}_{a}	1.190^{*}_{b}
3	0.991^{*}_{a}	-0.043^{*}_{a}	0.949^{*}_{a}
4	0.993^{*}_{a}	0.127^{*}_{a}	1.120^{*}_{b}
5	0.992^{*}_{a}	0.000_{a}	0.992^{*}
Panel D: Revenue spatial cumulative expansion-path scale economies			
Horizon, ϕ	$EPSE^{Dir}_{i\phi}(\boldsymbol{r})$	$EPSE^{Ind}_{In,i\phi}(\boldsymbol{r}) = EPSE^{Ind}_{Out,i\phi}(\boldsymbol{r})$	$EPSE^{Tot}_{In,i\phi}(\boldsymbol{r}) = EPSE^{Tot}_{Out,i\phi}(\boldsymbol{r})$
0	0.998^{*}	0.980^{*}	0.994^{*}_{a}
1	0.991^{*}_{a}	0.989^{*}_{a}	1.002^{*}
2	0.999^{*}	0.985^{*}	0.998^{*}
3	0.998^{*}	0.993^{*}	0.995^{*}_{a}
4	0.997^{*}	0.969^{*}_{a}	0.995^{*}_{a}
5	0.990^{*}_{a}	1.001^{*}	0.998^{*}

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

see that all the $EPSE^{Ind}_{i\phi}(c)$ and $EPSE^{Ind}_{i\phi}(r)$ are positive. We draw attention to two further differences between the indirect cost (revenue) RSE and EPSE. First, whereas the magnitudes of the direct cost (revenue) RSE and EPSE are always not markedly different, we observe big differences between the magnitudes of the indirect cost (revenue) RSE and EPSE. For example, for horizon 5, the indirect revenue RSE and EPSE are not significantly different from 0 and 1, respectively. Second, whereas we noted above that the magnitudes of the direct cost (revenue) RSE and EPSE are very persistent in future in-sample periods, we find that this is only the case for the indirect cost (revenue) EPSE. These two differences between the cost (revenue) indirect RSE and EPSE suggest that movements along an output (input) radial ray and output (input) expansion-path lead to far less

disparity between the direct cost (revenue) RSE and EPSE than we observe in the indirect case. Thus, the impact of allowing for the possibility that a bank may not lie on a radial ray is much bigger when we account for spillovers between banking networks which strengthens the above case for EPSE over RSE.

We now turn to discuss the total returns to scale results. Specifically, when there is an increase in the sample average bank's outputs (inputs) in the current period, the $RSE^{Tot}_{i\phi}$ and $EPSE^{Tot}_{i\phi}$ in Table 6 measure the percentage changes in this hypothetical bank's two symmetric total cost (revenue) measures in horizons 0–5. We note two things about these total returns. First, they incorporate both direct and indirect scale economies and therefore provide the overall picture of the optimal size of a bank. Second, as indirect spill-in and spill-out returns for the sample average bank are symmetric, this is also the case for the two total returns to scale measures. Along the same lines as the above discussion of the indirect RSE and EPSE, the case for EPSE over RSE is further reinforced when we consider the total returns. This is because in a large proportion of cases in Table 6 the magnitude of the difference between the total cost (revenue) RSE and EPSE is marked. We therefore focus on the total EPSE results. In contrast, Glass, Kenjegaliev, and Kenjegalieva (2020) did not indicate a preference between the total RSE and EPSE measures. This is because for their sample average U.S. bank both these measures are not statistically different from 1.

For the purposes of comparison, first recall that when we considered banks in isolation (i.e., when we overlooked indirect returns to scale spillovers between banking networks), we concluded from the $EPSE^{Dir}_{i\phi}(c)$ and $EPSE^{Dir}_{i\phi}(r)$ that the size of the sample average bank was approximately optimal. The issue is whether we reach the same conclusion when we account for indirect interbank returns to scale spillovers by considering $EPSE^{Tot}_{i\phi}(c)$ and $EPSE^{Tot}_{i\phi}(r)$. We can see from Table 6 that the magnitudes of the $EPSE^{Tot}_{i\phi}(c)$ ($EPSE^{Tot}_{i\phi}(r)$) and its direct counterpart are very similar. Therefore, from the magnitudes of these total EPSE, we once again conclude for horizons 0–5 that the size of the sample average bank is approximately optimal.

5.4 | Network perspectives of the scale economies of the G-SIBs and CCAR banks

Given our above preference for EPSE measures over RSE, in Table 7 for G-SIBs and banks included in the CCAR, we present the mean direct and total EPSE for horizons 0–3.[xvi] Note that we report two total EPSE, which is because for individual banks (or any other subset of the sample) the two total EPSE are asymmetric. As we previously noted, there is negligible feedback within the contemporaneous and dynamic direct parameters, so the direct EPSE in Table 7 are essentially own returns and are interpreted in the same way as the above direct-own returns for the sample average bank. The total EPSE in this table represent two network perspectives of a bank's returns to scale. The network perspective of $EPSE^{Tot}_{In,i\phi}$ is what a bank would focus on as this measure accounts for a bank's direct-own returns and the indirect *spill-in* returns to the bank. $EPSE^{Tot}_{In,i\phi}$ would also be of interest to bank regulators and antitrust policymakers who may also find the network perspective of $EPSE^{Tot}_{Out,i\phi}$ informative, as the latter accounts for a bank's direct-own returns and the indirect *spill-out* returns to other banks. $EPSE^{Tot}_{Out,i\phi}$ therefore indicates whether a bank's returns are optimal from the perspective of the cost or revenue implications for all the banks in the sample. Hence, when there is an increase in a bank's outputs (inputs) in the current period, the $EPSE^{Tot}_{In,i\phi}$ and $EPSE^{Tot}_{Out,i\phi}$ in Table 7 measure the percentage changes in the bank's two total cost (revenue) measures in horizons 0–3. These two total cost (revenue) measures comprise the bank's direct-own cost (revenue) and the asymmetric indirect costs (revenues) that spill-in and spill-out to and from the bank.

The capability of a G-SIB (CCAR bank) to influence the global banking system (domestic banking industry) will be affected by, among other things, the size of the bank and, via its interconnectedness, how its size affects, and is affected by, other banks. This highlights the importance of the optimal sizes of these banks and how this is affected by indirect returns to scale spill-outs and spill-ins, which we assess by comparing a bank's $EPSE^{Dir}_{i\phi}$ with its $EPSE^{Tot}_{Out,i\phi}$ and $EPSE^{Tot}_{In,i\phi}$. We can see for many of the banks in Table 7 that the contemporaneous direct-own cost and revenue EPSE are around 1. This indicates that when we overlook interbank geographical spillovers and consider banks in isolation, on average for the current period $t = 1, ..., T$, many of the banks in this table are around their optimal sizes. In line with the findings of Wheelock and Wilson (2018), these results suggest that non-negligible size caps would lead to such banks being sub-optimally small (large) from a cost (revenue) perspective. For the relatively small number of remaining banks in Table 7, the contemporaneous direct-own cost and revenue EPSE may not be considered to be around 1. These are all cases where the contemporaneous direct-own EPSE is marginally below 1 for cost (BNY Mellon, Bancorp and Zions) and / or revenue (BNY Mellon and Fifth Third). Again this suggests that size caps would lead to these banks being pushed further away from their minimum cost efficient scales.

TABLE 7 Drect and total expansion-path scale economies for G-SIBs and CCAR banks

	Contemporaneous horizon $\phi = 0$			Dynamic horizon $\phi = 1$			Dynamic horizon $\phi = 2$			Dynamic horizon $\phi = 3$		
	$EPSE^{Dir}_{i\phi}$	$EPSE^{Tot}_{In,i\phi}$	$EPSE^{Tot}_{Out,i\phi}$	$EPSE^{Dir}_{i\phi}$	$EPSE^{Tot}_{In,i\phi}$	$EPSE^{Tot}_{Out,i\phi}$	$EPSE^{Dir}_{i\phi}$	$EPSE^{Tot}_{In,i\phi}$	$EPSE^{Tot}_{Out,i\phi}$	$EPSE^{Dir}_{i\phi}$	$EPSE^{Tot}_{In,i\phi}$	$EPSE^{Tot}_{Out,i\phi}$
Dynamic spatial cost model												
G-SIBs												
JPMorgan Chase	1.000	1.035_b	1.126_b	0.998	1.039_b	0.997	0.997	1.041_b	0.867_a	0.996_a	1.042_b	0.842_a
Citibank	1.003_b	1.019_b	0.985	1.003	1.020_b	0.917_a	1.003	1.021_b	0.874_a	1.003	1.021_b	0.925_a
Bank of America	1.002	1.021_b	1.059_b	1.001	1.023_b	1.011	1.000	1.024_b	0.954	1.000	1.025_b	0.893_a
Wells Fargo	1.007_b	1.036_b	1.091_b	1.004_b	1.040_b	1.119_b	1.003	1.041_b	1.091	1.002	1.042_b	1.081
BNY Mellon	0.970_a	1.091_b	0.965_a	0.970_a	1.114_b	0.964_a	0.970_a	1.123_b	0.963_a	0.970_a	1.128_b	0.963_a
CCAR banks												
BBVA USA	0.985_a	1.002	1.044	0.985_a	1.005_b	1.028	0.985_a	1.007_b	1.030_b	0.985_a	1.007	1.036_b
Comerica	0.999	1.021_b	0.992	0.999	1.025_b	0.992	0.999	1.026_b	0.991	0.999	1.027_b	0.991
Fifth Third	0.984_a	1.015_b	1.042_b	0.984_a	1.021_b	0.943	0.984_a	1.023_b	0.957_a	0.984_a	1.024_b	0.937_a
Huntington	1.003	1.017_b	1.097_b	1.003	1.019_b	1.070	1.003	1.020_b	1.017	1.003	1.020_b	0.975_a
PNC Bank	0.996_a	1.018_b	1.048_b	0.995_a	1.021_b	1.024	0.994_a	1.023_b	0.931_a	0.994_a	1.023_b	0.994
Regions	0.998	1.008_b	1.040_b	0.998	1.009_b	0.996	0.998	1.010	0.934_a	0.998	1.010_b	0.965_a
Bancorp	0.973_a	1.003	0.977_a	0.973_a	1.012	0.980_a	0.973_a	1.017	0.981_a	0.973_a	1.019	0.982_a
Union	0.999	1.013_b	1.000	0.999	1.016_b	0.999	0.999	1.017_b	0.999	0.999	1.018_b	0.999
Zions	0.973_a	1.007	0.984_a	0.973_a	1.017_b	0.991	0.973_a	1.021_b	0.993	0.973_a	1.024_b	0.994
Dynamic spatial revenue model												
G-SIBs												
JPMorgan Chase	0.996_a	0.986_a	1.021_b	0.995_a	0.989_a	0.987_a	0.996_a	0.987_a	1.062_b	0.995_a	0.988_a	0.901
Citibank	1.000	0.994_a	1.026_b	0.999	0.996	0.967_a	1.000	0.994_a	1.038_b	0.999	0.995_a	1.041_b
Bank of America	0.997	0.987_a	1.022_b	0.996_a	0.990_a	0.999	0.997	0.988_a	1.055_b	0.996_a	0.989_a	1.025
Wells Fargo	0.994_a	0.986_a	1.013_b	0.993_a	0.989_a	1.080_b	0.993_a	0.987_a	1.043_b	0.993_a	0.988_a	1.006_b
BNY Mellon	0.987_a	0.965_a	0.968_a	0.967_a	0.966_a	0.968_a	0.967_a	0.966_a	0.968_a	0.967_a	0.966_a	0.968_a
CCAR banks												
BBVA USA	0.996_a	0.991_a	1.006_b	0.996_a	0.993	1.036_b	0.996_a	0.991_a	1.008_b	0.996_a	0.992_a	1.012_b
Comerica	0.994_a	0.988_a	0.996_a	0.994_a	0.990	0.996	0.994_a	0.989_a	0.996	0.994_a	0.989_a	0.996_a
Fifth Third	0.974_a	0.974_a	0.988_a	0.974_a	0.975_a	1.003	0.973_a	0.974_a	0.989_a	0.976_a	0.975_a	0.989
Huntington	0.992_a	0.988_a	0.997	0.992_a	0.989_a	0.988_a	0.992_a	0.988_a	0.996_a	0.992_a	0.989_a	0.998
PNC Bank	0.994_a	0.986_a	1.008_b	0.993_a	0.988_a	0.983_a	0.993_a	0.987_a	1.012_b	0.993_a	0.988_a	1.037_b

(Continues)

TABLE 7 (Continued)

	Contemporaneous horizon $\phi = 0$			Dynamic horizon $\phi = 1$			Dynamic horizon $\phi = 2$			Dynamic horizon $\phi = 3$		
	$EPSE^{Dir}_{i\phi}$	$EPSE^{Tot}_{In,i\phi}$	$EPSE^{Tot}_{Out,i\phi}$	$EPSE^{Dir}_{i\phi}$	$EPSE^{Tot}_{In,i\phi}$	$EPSE^{Tot}_{Out,i\phi}$	$EPSE^{Dir}_{i\phi}$	$EPSE^{Tot}_{In,i\phi}$	$EPSE^{Tot}_{Out,i\phi}$	$EPSE^{Dir}_{i\phi}$	$EPSE^{Tot}_{In,i\phi}$	$EPSE^{Tot}_{Out,i\phi}$
Regions	0.996_a	0.990_a	1.009_b	0.996_a	0.992	1.011_b	0.996_a	0.990_a	1.013_b	0.996_a	0.991_a	1.019_b
Bancorp	0.990_a	0.989_a	0.996_a	0.990_a	0.990_a	0.993_a	0.990_a	0.989_a	0.995_a	0.990_a	0.990_a	0.994_a
Union	0.997	0.992_a	1.001	0.997	0.993_a	1.002	0.997	0.992_a	1.002	0.997	0.992_a	1.002
Zions	0.991_a	0.987_a	0.991_a	0.991_a	0.988_a	0.991_a	0.991_a	0.988_a	0.991_a	0.991_a	0.988_a	0.991_a

Note: G-SIBs denotes the global systemically important banks. CCAR banks are those included in the *Comprehensive Capital Analysis and Review* and comprise the G-SIBs and the additional above banks. a and b denote returns that are significantly less than or greater than 1, respectively.

Our conclusions from the dynamic direct-own EPSE in Table 7 about the (sub-)optimal sizes of the banks in each of the three future in-sample periods are the same as those above for the current period from the contemporaneous direct-own EPSE. This is because the magnitudes of the dynamic direct-own cost and revenue EPSE are very similar to the corresponding contemporaneous estimates. As a result, our findings from the dynamic direct-own EPSE reiterate our above conclusions from the corresponding contemporaneous estimates about the impact of size caps on these banks.

To assess the impact of the two network perspectives we now consider the $EPSE^{Tot}_{In,i\phi}$ and $EPSE^{Tot}_{Out,i\phi}$ in Table 7, as both these measures incorporate the direct-own returns and also indirect spill-in or spill-out returns. For a number of the banks in this table the contemporaneous and dynamic $EPSE^{Tot}_{In,i\phi}(c)$ are around 1 and are marginally higher than the corresponding $EPSE^{Dir}_{i\phi}(c)$. The clear exception where the contemporaneous and dynamic $EPSE^{Tot}_{In,i\phi}(c)$ are markedly greater than 1 and larger than the corresponding $EPSE^{Dir}_{i\phi}(c)$ is BNY Mellon, and other exceptions where to a lesser extent this is the case include JPMorgan Chase and Wells Fargo. These contemporaneous and dynamic $EPSE^{Tot}_{In,i\phi}(c)$ suggest that the three banks are sub-optimally large, which is not out of line with what the corresponding revenue estimates suggest. However, the suggestion that these banks are sub-optimally large is at odds with the contemporaneous and dynamic $EPSE^{Dir}_{i\phi}(c)$ for BNY Mellon and JPMorgan Chase and some of the dynamic $EPSE^{Dir}_{i\phi}(c)$ results for Wells Fargo. Focusing for the moment on only the contemporaneous and dynamic $EPSE^{Tot}_{In,i\phi}$ results suggests that appropriate size caps would move these three banks in the direction of their minimum efficient scales. This conclusion though may be viewed by regulators and anti-trust policymakers as representing half the picture because it overlooks the implications of the $EPSE^{Tot}_{Out,i\phi}$ results, which we illustrate using the $EPSE^{Tot}_{Out,i\phi}(c)$ estimates for JPMorgan Chase.

We can see from Table 7 that the contemporaneous $EPSE^{Tot}_{Out,i\phi}(c)$ for JPMorgan Chase is well above 1, which is consistent with the corresponding $EPSE^{Tot}_{In,i\phi}(c)$ being greater than 1, and suggests that the bank is sub-optimally large. However, in horizon 1 the $EPSE^{Tot}_{Out,i\phi}(c)$ for JPMorgan Chase is approximately equal to 1 and in horizons 2 and 3 is well below 1 (e.g., 0.842 in horizon 3). This suggests that although an appropriate size cap on JPMorgan Chase in the current period would push the bank towards its contemporaneous minimum efficient scale, this would be sub-optimal in a dynamic setting. This conclusion and other similar findings from our analysis suggest that dynamic spatial measures, such as $EPSE^{Tot}_{Out,i\phi}$ and $EPSE^{Tot}_{In,i\phi}$, can provide additional insights

for the stakeholder policy decisions of bank regulators and antitrust policymakers.

As we noted above, the network perspective of $EPSE^{Tot}_{In,i\phi}$ is the reason why a bank would focus on this measure. For all but one bank in Table 7 the magnitudes of the contemporaneous and dynamic $EPSE^{Tot}_{In,i\phi}(c)$ are similar. For BNY Mellon, however, there is a noticeable difference between its contemporaneous $EPSE^{Tot}_{In,i\phi}(c)$ and its corresponding dynamic measures. When there are such differences between the contemporaneous and dynamic $EPSE^{Tot}_{In,i\phi}$, a bank will be faced with a situation whereby, if it acts on the estimate of its contemporaneous $EPSE^{Tot}_{In,i\phi}$ by, for example, reducing its size by a certain percentage in the current period, this can lead to sub-optimal dynamic returns in the following periods. To manage these dynamic returns to scale effects, we suggest that a bank should aim to optimize its contemporaneous and dynamic returns over the time frame of its future plans, which would involve some returns being sub-optimal for particular periods within this time frame.

6 | SUMMARY AND CONCLUSIONS

This paper makes two contributions to the literature. The first contribution extends the methods for static external returns to scale in Glass, Kenjegaliev, and Kenjegalieva (2020) to measure the dynamic persistence in future in-sample periods of: (i) returns to scale that are internal to a bank; and (ii) external returns to scale. These new methods enable us to assess whether the classifications of contemporaneous internal and external returns to scale are consistent with the classifications of the dynamic internal and external returns in future in-sample periods. These measures of the persistence of internal and external returns can be therefore be used to assess the dynamic optimality of a change in the size of a bank.

A priori, an EPSE measure can be viewed as more appropriate than the corresponding RSE. This is because, in contrast to the latter, the former allows for the possibility that a bank may not lie on a radial ray. For the sample average bank we find that the magnitudes of the direct-own cost (revenue) RSE and EPSE are not markedly different, whereas when we account for indirect spillover returns we observe non-negligible differences between the cost (revenue) oriented total RSE and EPSE. This raises the question: which of these two total returns to scale measures do we prefer? Glass, Kenjegaliev, and Kenjegalieva (2020) did not indicate a preference between these measures because for their sample average U.S. bank both measures are not statistically different from 1. Given the non-negligible differences we observe between the cost (revenue) oriented total RSE and EPSE, we conclude that accounting for spillovers strengthens the above case for EPSE over RSE and explains why we focus on EPSE. Moreover, for the sample average bank we find that the direct-own, indirect and total EPSE are very persistent. This is based on these returns being non-negligible and significant for at least 5 years.

The focus in Glass, Kenjegaliev, and Kenjegalieva (2020) is the presentation of their methods for static internal and external returns to scale and they only provide a small and general demonstration of these methods for the costs of large and medium-sized U.S. banks. In light of this, the second contribution of our paper is to carry out a policy focused empirical analysis that directly relates to size caps by applying the methods we introduce to the costs and revenues of key large U.S. banks; that is, 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).

To assess how accounting for returns to scale spillovers impacts the conclusions about the optimality of the sizes of G-SIBs and CCAR banks, we compare a bank's direct-own EPSE with its $EPSE^{Tot}_{In,i\phi}$ and $EPSE^{Tot}_{Out,i\phi}$. These two total EPSE measures represent two network perspectives of a bank's returns to scale. The network perspective of $EPSE^{Tot}_{In,i\phi}$ is what a bank would focus on as this measure accounts for a bank's direct-own returns and the indirect returns that spill-in to the bank. $EPSE^{Tot}_{In,i\phi}$ would also be of interest to bank regulators and antitrust policymakers who may also find the network perspective of $EPSE^{Tot}_{Out,i\phi}$ informative. This is because $EPSE^{Tot}_{Out,i\phi}$ accounts for a bank's direct-own returns and its indirect spill-out returns to other banks. $EPSE^{Tot}_{Out,i\phi}$ therefore indicates whether a bank's returns are optimal from the perspective of the cost (revenue) implications for all the banks in the analysis.

To demonstrate the implications of the contemporaneous and dynamic $EPSE^{Tot}_{In,i\phi}$ measures for a bank, we focus on the results for a particular G-SIB, BNY Mellon. We observe a noticeable difference between its cost oriented contemporaneous and dynamic $EPSE^{Tot}_{In,i\phi}$ measures. When there is such a difference between these measures a bank will be faced with a situation whereby, if it acts on its contemporaneous $EPSE^{Tot}_{In,i\phi}$ by (not) changing its size in the current period, this can lead to sub-optimal dynamic returns in the following periods. In this situation, we suggest that a bank seeks to manage these dynamic returns to scale effects by optimizing its contemporaneous and dynamic returns over the time frame of its future plans, which would involve some returns being sub-optimal in particular periods within this time frame.

To demonstrate the implications of the contemporaneous and dynamic $EPSE^{Tot}_{Out,i\phi}$ measures for bank regulators, we focus on the results for another G-SIB, JPMorgan Chase. The cost oriented contemporaneous $EPSE^{Tot}_{Out,i\phi}$ measure suggests that an appropriate size cap would be consistent with JPMorgan Chase operating at its minimum efficient scale in the current period. However, its cost oriented dynamic $EPSE^{Tot}_{Out,i\phi}$ measures suggest that any size cap would lead to a sub-optimal dynamic scale in the following periods. This and other similar conclusions from our analysis suggest that dynamic spatial measures can provide additional insights for bank regulators to further inform their policymaking and its impacts on the various stakeholders.

Finally, we recognize that we focus on extending a common non-spatial approach to the modelling of the banking production technology (e.g., Wheelock & Wilson, 2012, 2018) to simultaneously account for the contemporaneous and dynamic spatial interactions among the banks. Due to our focus being on these spatial interactions we do not explicitly model two further aspects of the banking production technology that feature in related research: namely; the level of risk which is endogenously related to the bank business model (Delis et al., 2017), and the level of diversification of banking activities (Laeven & Levine, 2007). Rather we adopt the approach in Orea et al. (2018) and implicitly account for such factors by modelling the spatial dependencies among the banks, as this modelling approach represents a way to mitigate omitted variables and the resulting endogeneity. Given our paper sets out a dynamic spatial framework to model the banking production technology, our framework can be used in future research that focuses on explicitly modelling these two further factors.

ACKNOWLEDGEMENTS

The authors would like to thank the handling editor for processing our paper and three anonymous reviewers for their comments.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

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

ENDNOTES

[i] We recognize that size caps might be imposed for other reasons, for example, to reduce the risk exposure of the largest banks. A full assessment of size caps should therefore take account of all the impacts of size caps. Such an assessment is outside the scope of this paper and we instead focus on the returns to scale implications of size caps.

[ii] Note that the profit function that is often estimated in the banking literature is the alternative functional form, and not its standard counterpart. This is because profit in the alternative specification is a function of, among other things, outputs, where the impacts of these quantities accounts for higher quality outputs via the additional revenue they generate.

[iii] The alternative profit function, of course, takes into account costs and revenues. We explored using a dynamic spatial alternative profit function for large U.S. banks to calculate the contemporaneous and dynamic internal and external returns to scale. We did not pursue this alternative function because we found that the coefficient on the time lag of the spatially lagged profit variable (which is key in the calculations of the dynamic external returns to scale) was not significant. From the positive and negative significant coefficients we observe on the time lags of the spatially lagged cost and revenue variables, it is evident that is because these two effects offset one another in the dynamic spatial alternative profit model. By focusing on the dynamic spatial cost and revenue models, we get insights into how these offsetting effects impact the dynamic external (cost and revenue) returns to scale.

[iv] We thank an anonymous reviewer for making the point that in this situation the regulator would be faced with prioritizing particular stakeholder interests.

[v] We use balanced panel data because the asymptotic properties of spatial panel data estimators breakdown when the panel is unbalanced and the reason why data are missing is not known (Elhorst, 2009).

[vi] Note that there are two contemporaneous SAR variables in the Glass et al. (2020a, 2020b) models, but only one in Equation (1). This is because via multiple spatial weights matrices their static spatial models account for multiple contemporaneous spatial regimes. This is something we overlook to simplify matters, as the focus of our study is the complexity of dynamic spatial interactions.

[vii] In line with the properties of the estimator, N is large in the empirical analysis (201). In panel data T is often small, and so in a dynamic setting the time lag of the dependent variable is endogenous. When this is the case in a dynamic spatial setting, the time lag of the SAR variable will also be endogenous. In the empirical analysis T is not particularly small (22 years), so following Yu et al. (2008), the time lag of the SAR variable is taken to be exogenous.

[viii] Following LeSage and Pace (2009) in the spatial econometrics literature, we obtain the standard errors for the estimates of the parameters in Equation (1) using a mixed analytical-numerical Hessian matrix. When the equation for a second order derivative of the log-likelihood function in this matrix does not contain the spatial multiplier matrix $(\mathbf{U}^{-1} = (\mathbf{I}_N - \delta\mathbf{W})^{-1}$ period, which plays a key role in the reduced form of Equation (1), see Equation (2), as it captures the global spatial interactions between a bank and its 1st order, 2nd order, etc. neighbouring banks) the derivative is computed analytically, otherwise it is computed numerically. Evaluating such derivatives analytically is less sensitive to badly scaled data, while numerical evaluation when N is very large avoids any difficulties (or lengthy

computation time) associated with the evaluation of the spatial multiplier matrix. We do not adjust the standard errors for clustering on the basis that '...it is difficult to motivate clustering if the regression function already includes fixed effects' (Abadie et al., 2017, p. 1). This is particularly so for clustering at the bank level due to the bank level fixed effects in Equation (1). Moreover, along the same lines, we posit that the spatial bank interdependencies that Equation (1) accounts for make it difficult to motivate clustering at a geographical level.

[ix] These 12 states are as follows: Colorado; Georgia; Hawaii; Illinois; Kansas; Kentucky; Louisiana; Minnesota; Missouri; New Hampshire; Tennessee; and Wisconsin.

[x] Two further important aspects of the banking production technology that feature in related research which an anonymous reviewer highlighted are the level of risk which is endogenously related to the bank business model (Delis et al., 2017), and the level of diversification (Laeven & Levine, 2007). Given our focus is on simultaneously accounting for contemporaneous and dynamic spatial interactions among banks, we do not explicitly model these factors. Rather, and as we noted in Section 4.1, we follow the approach of Orea et al. (2018) and implicitly account for such factors by modelling the spatial dependencies among the banks, as this modelling approach represents a way to mitigate omitted variables and the resulting endogeneity. Given our paper sets out a dynamic spatial framework to model the banking production technology our framework can be used in future research that focuses on explicitly modelling these two factors.

[xi] Recall from the above discussion of Equation (1) that exogenous spatial weights is an assumption of the modelling. As the off-diagonal elements of the average matrix from (3) are based on state level micro information on branch locations, and the dependent variables are at the aggregate bank level, it is reasonable to take the spatial weights to be exogenous.

[xii] Frequently in the spatial literature, the weights matrix is normalized by its row sums. This is suited to binary spatial weights that represent, for example, contiguous geographical areas, which is very different to the spatial arrangement of the branches in our application. If we row-normalized our non-binary weights we would lose the information on the relative branch network intensities.

[xiii] Note that there are two $\mathbf{W}r_t$ parameters in the Glass et al. (2020b) static spatial revenue model. This is because their model allows for multiple contemporaneous spatial weights matrices (i.e., multiple spatial regimes). This is something we overlook to simplify matters as the focus of our study is the complexity of dynamic spatial interactions.

[xiv] For brevity we do not report these standard non-spatial models, but they are available from the corresponding author on request.

[xv] The exception that is not significant is the estimate in Table 4 of the q_3 (non-interest income) direct elasticity for horizon 21.

[xvi] The group of banks that the CCAR covers comprise the G-SIBs and a further group. The CCAR has been conducted annually since 2011 and from 2013 has included results for individual banks. The G-SIBs have been ever present in the CCAR, while the other banks in Table 7 have featured in the CCAR over the following periods: BBVA USA (2013–2017); Comerica (2013–2016); Fifth Third (2011–2019); Huntington (2014–2017); PNC Bank (2011–2019); Regions (2011–2018); Bancorp (2011–2019); Union (2014); Zions (2013–2016).

REFERENCES

Abadie, A., Athey, S., Imbens, G. W., & Wooldridge, J. (2017). When should you adjust standard errors for clustering? NBER Working Paper No. 24003.

Anselin, L. (1988). *Spatial econometrics: Methods and models*. 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., 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., & 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.

Ciccarelli, C., & Elhorst, J. P. (2018). A dynamic spatial econometric diffusion model with common factors: The rise and spread of cigarette consumption in Italy. *Regional Science and Urban Economics*, *72*, 131–142.

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

Delis, M., Iosifidi, M., & Tsionas, M. G. (2017). Endogenous bank risk and efficiency. *European Journal of Operational Research*, *260*, 376–387.

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

Elhorst, J. P. (2009). Spatial panel data models. In M. M. Fischer & A. Getis (Eds.), *The handbook of applied spatial analysis*. Springer.

Federal Reserve Board. (2019). Comprehensive capital analysis and ReviewQuantitative results, 2013–2019. https://www.federalreserve.gov/supervisionreg/files/public_results_CCAR.csv

Feng, G., & Serletis, A. (2010). Efficiency, technical change, and returns to scale in large US banks: Panel data evidence from an output distance function satisfying theoretical regularity. *Journal of Banking and Finance*, *34*, 127–138.

Feng, G., & Zhang, X. (2014). Returns to scale at large banks in the US: A random coefficient stochastic frontier approach. *Journal of Banking and Finance*, *39*, 135–145.

Financial Stability Board. (2019). *2019 list of global systemically important banks* (*G-SIBs*). https://www.fsb.org/wp-content/uploads/P221119-1.pdf

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.

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

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

Hughes, J. P., & Mester, L. J. (2013). Who said large banks don't experience scale economies? Evidence from a risk-return-driven cost function. *Journal of Financial Intermediation*, *22*, 559–585.

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.

Kovner, A., Vickery, J., & Zhou, L. (2014). Do big banks have lower operating costs? *Federal Reserve Bank of New York Policy Review*, *20*, 1–27.

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*. CRC Press.

LeSage, J. P., & Sheng, Y. (2014). A spatial econometric panel data examination of endogenous versus exogenous interaction in Chinese province-level patenting. *Journal of Geographical Systems*, *16*, 233–262.

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

Noulas, A. G., Ray, S. C., & Miller, S. M. (1990). Returns to scale and input substitution for large US banks. *Journal of Money, Credit and Banking*, *22*, 94–108.

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.

Restrepo-Tobón, D., & Kumbhakar, S. C. (2015). Nonparametric estimation of returns to scale using input distance functions: An application to large US banks. *Empirical Economics*, *48*, 143–168.

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.

Stern, G. H., & Feldman, R. (2009). Addressing TBTF by shrinking financial institutions: An initial assessment. In *The region* (pp. 8–13). Federal Reserve Bank.

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 US banking. *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.

How to cite this article: Glass, A. J., & Kenjegalieva, K. (2023). Returns to scale, spillovers and persistence: A network perspective of U.S. bank size. *International Journal of Finance & Economics*, 1–28. https://doi.org/10.1002/ijfe.2776