Spatial Scale and Product Mix Economies in U.S. Banking with Simultaneous Spillover Regimes

Anthony J. Glass*, Amangeldi Kenjegaliev†,‡ and Karligash Kenjegalieva§

November 2019

Abstract

The literature on bank scale economies focuses on the familiar type of returns to scale that are internal to the firm. Using a spatial approach, we analyze returns to scale for banks that are made up of external (i.e., spillover) economies. We extend ray-scale economies (RSE), expansion-path scale economies (EPSE) and expansion-path subadditivity (EPSU) to the spatial case. This involves introducing direct and composite and decomposed indirect RSE, EPSE and EPSU. These direct and indirect measures relate to the cost implications for a firm from a change in: (i) the firm's output levels that are, as is standard, under its control; and (ii) the composite/decomposed spillover effect on the firm's output levels, which is primarily, but not entirely, outside its control. We include an application to U.S. banks $(1998-2015)$ that allows a bank to simultaneously belong to a number of spatial networks, which is typically what we observe for firms. For large banks we find constant direct RSE and EPSE, and zero composite indirect RSE and constant composite indirect EPSE. These composite indirect results do not counteract any policy suggestions from the direct RSE and EPSE concerning the debate on whether there should be size caps on very large U.S. banks. The direct RSE and EPSE for large banks suggest that these banks use society's resources efficiently to provide their services. Size caps on very large banks would place downward pressure on these direct RSE and EPSE results, which could lead to large banks using society's resources inefficiently.

Key words: Productivity and competitiveness; Internal and external returns to scale; Spatial cost function; Branch networks.

[*] School of Business and Economics and Centre for Productivity and Performance, Loughborough University, Loughborough, Leics, UK, LE11 3TU. Email: a.j.glass@lboro.ac.uk

[†] Hull University Business School, Hull, HU6 7RX, UK. Email: A.Kenjegaliev@hull.ac.uk

[‡] Corresponding author

[§] School of Business and Economics and Centre for Productivity and Performance, Loughborough University, Loughborough, Leics, UK, LE11 3TU. Email: k.a.kenjegalieva@lboro.ac.uk

1 Introduction

The "too-big-to-fail" (TBTF) status of very large U.S. banks played a key role in the 2008 financial crisis as it promoted excessive risk taking. The 2010 Dodd-Frank reforms to guard against a repeat of this risk taking involved: tightening the regulatory regime through, for example, more stringent bank liquidity constraints; and establishing a formal process to resolve large bank failures with the intention of ensuring that no bank is TBTF. Fisher and Rosenblum (2012), among others, argue that Dodd-Frank could have gone further to prevent TBTF banks by introducing bank size caps. Interestingly though many of the largest banks are much bigger now than they were before the crisis (Wheelock and Wilson, 2018).

From a cost perspective, the categorization of returns to scale (increasing/constant/decreasing) determines whether size caps have a negative or positive effect on how efficiently society's resources are used to provide banking services (Stern and Feldman, 2009; Wheelock and Wilson, 2018). This highlights the importance of measuring returns to scale in banking, which is particularly so for U.S. banks as all three returns to scale categories have been reported in the literature. Despite such evidence for U.S. banks there is some consensus across a number of influential studies (Wheelock and Wilson, 2009; 2012; 2018; Hughes and Mester, 2013; Kovner *et al.*, 2014; Mester, 2010), which primarily point to non-negligible increasing returns, although in some of these studies there is also a small amount of evidence of constant returns. We do not, however, review the large literature on scale economies in U.S. banking as this literature focuses on the familiar scale economies that are internal to the firm. Instead we pursue a new line of inquiry on bank scale economies by using a spatial modeling approach to analyze external returns to scale in U.S. banking.

External returns to scale refer to the economies that benefit a firm because of the way in which the industry it operates in is organized (Economist, 2008). External returns are not therefore as tightly defined as internal returns to scale. Consequently, in different literatures external returns to scale are defined more specifically. For example, external economies are a prominent feature of the urban economics literature, where these returns are referred to as agglomeration economies. These economies represent the benefits from firms and people locating near one another in cities and industrial clusters. It is still the case though that agglomeration economies can take a range of forms so from an empirical perspective it is necessary to focus on a particular type of benefit, which traditionally involves estimating a wage equation. We, on the other hand, consider external economies using an OR production framework by firmly embedding these economies within the relevant function (cost, revenue, etc.).

Following the large literature that uses non-spatial methods to estimate the familiar scale economies that are internal to a bank would involve overlooking potentially important external scale economies. These external economies are important because, like internal scale economies, they can be related to a firm's outputs. In particular, internal economies relate to a firm's output levels that are within its control, while external economies relate to an individual firm's output levels that are due to spillovers to the firm from other firms, which are primarily, but not entirely, outside the firm's control. These spillovers are likely to depend on some industry level factor such as the size or structure of the industry, which an individual firm may have a degree of control over through its involvement in the interrelated strategic choices of firms.[1] The upshot is that the classification of both internal and external economies (increasing/constant/decreasing) is important because if they have the same classification

[1] We thank an anonymous reviewer for highlighting that external economies may not be entirely outside an individual firm's control.

they will have the same type of implications for costs. If they are classified differently the implications for costs will differ and the issue is whether the effect of the internal economies is greater than that of the external economies or vice-versa. We illustrate this point in detail when discussing the policy implications from our application in subsection 4.3.

The external scale economies we analyze are the result of the spatial correlation between banks' internal scale economies. The mechanism in our analysis that leads to these external economies is as follows. Using a spatial cost function we estimate inter-bank spillovers, which then allows us to analyze external returns to scale by measuring the spillover to a bank's cost due to the spillover effects on its outputs. These spillover effects on its outputs arise from a change in other banks' outputs, which represents the attribution to the outputs of the spatial correlation between banks' costs and the spatial correlations between the effects of a bank's outputs on its costs. These spatial correlations are the result of banks operating in the same markets, where a banking market is typically taken to be a metropolitan statistical area (MSA) or non-MSA county (Hirtle, 2007). The spatial correlations can be negative or positive and the external scale economies are based on the net spatial correlation. Although we do not use structural economic theory to model the channels that lead to the spatial correlations, which is an area we leave for further work, negative spatial correlation in the spatial literature is attributed to the effects of competition (e.g., Kao and Bera, 2013, Boarnet and Glazer, 2002, and Garrett and Marsh, 2002). Positive spatial correlation is typically associated with common phenomena between neighbors such as market growth and headline changes in local and regional economies.

Our spatial cost function is made up of the bank's technology, which is a relationship between the bank's cost and its outputs and input prices, and spatial variables that shift this cost technology. These spatial variables are spatial lags of the bank's cost, output and input price variables. In contrast to how scale economies are calculated from a non-spatial function, it is not simply a matter of using fitted coefficients from our spatial cost function to calculate the spatial scale economies we propose. This is because, as is well-established in the spatial literature, any coefficient from for simplicity a log-linear cost function augmented with a spatial lag of the dependent variable is not an elasticity. To illustrate, for one of a firm's input prices or outputs the elasticity is instead a function of the coefficient on the variable and also the coefficient on the spatial lag of the dependent variable. To see this note that when one of a firm's own variables changes there are two effects. First and as is standard in non-spatial models, this change will affect the firm's cost, which is what the coefficient on the firm's own variable measures. Second, this change will affect the cost of each of the other firms in the sample via the spatial lag of the dependent variable and this effect will partially reverberate back to the firm whose own variable changed, which is referred to as feedback.[2] By combining these two effects we obtain what LeSage and Pace (2009) (LSP) refer to as the direct elasticity of a firm's variable.

LSP also derive from a model with a spatial lag of the dependent variable indirect and total elasticities, which we also compute. An indirect elasticity can be calculated in two ways depending on whether the focus is on spill-ins or spill-outs. We are interested in returns to scale spill-ins to a firm so the indirect elasticity is the effect on a firm's dependent variable from the spillover impact that permeates to its independent variable. We also depart slightly from the terminology in the spatial literature by referring to this elasticity as a composite indirect measure to distinguish it from its three

[2] The feedback is a partial rebound effect because spatial models are based on the assumption that there is a fading memory across space, which we revisit further in the paper and is entirely reasonable for many applications. Typically this feedback is small in empirical applications.

components, which we calculate to consider different spillover sources. By summing the direct and composite/decomposed indirect elasticities we obtain the total elasticity.

A spatial error model uses a spatial autocorrelated error term to account for global cross-sectional nuisance dependence, where global spatial dependence refers to 1st order neighbor and higher order neighbor spatial interactions, and local spatial dependence is confined to 1st order neighbor spatial interaction. We, on the other hand, as LSP recommend use a model specification that accounts for economically substantive cross-sectional global and local dependencies. This global dependence is accounted for using the spatial lag of the dependent variable and the local dependencies are captured using spatial lags of the independent variables. Important features of our model include the following. (a) By including the spatial lags of the dependent and independent variables we address a potential source of omitted variable bias and any associated parameter inconsistency, whereas the inclusion of the spatial autocorrelated error term in the spatial error model has no bearing on the consistency of the parameter estimates and only reduces their standard errors (i.e., the inclusion of this term only improves the efficiency of the parameter estimates). (b) Crucially, to enable us to calculate the spatial scale economies we propose and as we described above, since our model contains the spatial lag of the dependent variable we can relate the indirect elasticities to a firm's output(s). One cannot calculate such economies from a spatial error model because it is now well-known that the only type of indirect elasticity it yields relates to the disturbance.[3]

We focus on extending the methods to calculate non-spatial scale and product mix economies across the multiple products in banking (Berger *et al.*, 1987; Wheelock and Wilson, 2001) to the spatial case. It is useful to note that the origins of these non-spatial scale and product mix economies emanate from the cost or supply side benefits from joint production that Baumol *et al.* (1982) emphasize. Applying this idea to banking, Berger *et al.* (1987) posit that these benefits are in the form of two sources of cost savings from joint production. First, if excess capacity exists, joint production presents an opportunity to allocate fixed costs more widely by spreading fixed or quasi-fixed brick and mortar branch costs, or loan officer and teller expenses across a wider product mix. Second, joint production presents the opportunity to exploit information economies because information that is obtained from servicing a customer's deposits and/or loans can be reused. Reusing this information, for example, reduces the cost of evaluating the default probabilities for other types of loans.

Using the direct, composite and decomposed indirect and total parameters associated with a bank's independent variable, we set out the methodology to extend the non-spatial approaches to scale and product mix economies in banking to the spatial case. The direct (composite/decomposed) [total] measures of scale and product mix economies we introduce relate to the cost implications for a bank from a change in: (i) the bank's output levels that are, as is standard, under its control; (ii) (the composite/decomposed spillover effect on the bank's output levels from other banks, which are primarily, but not entirely, outside a bank's control); and (iii) [the bank's output levels that are under its control and the composite spillover effect on its output levels combined].[4] Direct scale and product mix economies are therefore akin to the standard internal measures from a non-spatial model and borrowing terminology from the spatial literature we label the external scale and product mix economies as indirect measures. Total scale and product mix economies represent the combined effect of the direct and composite/decomposed indirect measures.

[3] We thank an anonymous reviewer for suggesting that we highlight why the spatial economies we propose cannot be calculated from a spatial error model.

[4] Although we set out the direct, composite and decomposed indirect and total measures of scale and product mix economies for a cost technology, our methodology is in no way limited to a cost setting and can easily be adapted to other technologies (e.g., revenue and profit).

Putting the contributions of this paper into a wider methodological context, to the best of our knowledge, there are just three studies (Glass *et al.*, 2013; 2016; Glass and Kenjegalieva, 2019) that compute spatial returns to scale from a spatial functional form of production and cost technologies. For European countries, Glass *et al.* (2016) compute spatial ray-scale economies (RSE), which we discuss in due course, and Glass *et al.* (2013) propose a partial spatial decomposition of total factor productivity (TFP) growth. Their decomposition is partial as its omits the changes in technical and allocative efficiency spillovers. Glass and Kenjegalieva (2019) address these omissions and apply their methodology to spatially decompose TFP growth for U.S. banks. In both these spatial TFP growth papers spatial RSE is the basis for the calculation of the spatial returns to scale efficiency change component in the decompositions.

We build on the three aforementioned studies in the following four respects. First, in contrast to the first two papers, our paper is closely aligned to OR as we undertake a firm level analysis of spatial returns to scale. Second, although the third study also analyzes U.S. banks it computes the change in spatial returns to scale efficiency, whereas we report the level estimates of spatial returns to scale which enables us to focus on the minimum efficient size of a bank. Third, in addition to computing spatial RSE, we extend two further measures of scale and product mix economies (expansion-path scale economies ($EPSE$) and expansion-path subadditivity ($EPSU$)) to the spatial setting. Fourth, the spatial model we use to calculate the spatial scale and product mix economies is more technically advanced than the standard model in the above three studies, which has a single spatial lag of the dependent variable. Our model, on the other hand, includes multiple different specifications of the spatial lag of the dependent variable. This allows a firm to simultaneously belong to a number of spatial networks (otherwise known as spatial regimes in the literature), which is what we often observe in practice. Due to banks belonging to multiple spatial networks in our model, we obtain rich decomposed indirect RSE, $EPSE$ and $EPSU$ measures that relate to the spatial cost correlation between banks in the combined networks, and the spatial correlation between the effect of a bank's outputs on its cost in each network. Mathematically it is not possible to obtain decomposed indirect RSE, $EPSE$ and $EPSU$ measures that relate to the spatial cost correlation between banks in each network, which further in the paper we revisit in more detail.

Our banking application fits this multi-network setting very well because all the banks in our sample have two distinct branch networks. The first is a bank's full service brick and mortar branch network and the second is a bank's network of other types of full and limited service branches. Splitting up a bank's branches in this way makes sense as the spatial interaction between the banks in the two networks will differ because, as we discuss in more detail further in the paper, compared to brick and mortar branches there is a much smaller number of other branches as often the latter focus on centralized specialist activities (e.g., full service cyber offices, limited service loan production offices and limited service consumer credit offices).

In a two-dimensional output space composite indirect RSE refers to an equiproportional increase in a firm's two output levels attributable to spillovers along the radial ray. Composite indirect $EPSE$ relates to an incremental change in a firm's two outputs attributable to spillovers along its expansion-path for these types of outputs, which allows for the possibility that this path does not coincide with the composite indirect radial ray. Our empirical analysis of U.S. banks over the period $1998 - 2015$ often points to, on average, constant direct RSE and $EPSE$, whilst we also report zero composite indirect RSE and constant composite indirect $EPSE$. This suggests along the relevant radial ray and expansion-path that there is a much bigger difference between the expansion of a bank's outputs that are attributable to spillovers from other banks, than there is between the standard expansion of its outputs that are under its control. Greater similarity between direct RSE and $EPSE$ vis-à-vis the corresponding composite indirect measures is not surprising. This is because underlying direct/internal scale economies is the standard theoretical cost function that is monotonically increasing in a firm's output levels that are under its control, whereas there is no such theoretical relationship between a firm's cost and its output levels due to spillovers, which are primarily, but not entirely, outside its control.

From our key results on the spatial economies there are some policy implications on whether there should be caps on the size of very large banks, which we discuss in detail further in the paper. Although these policy implications are from a recent data sample, one should be cautious about how long these implications are valid for because the U.S. banking industry is known to change quickly. Also, given the novelty of our above findings and the large literature on bank efficiency covering a wide range of countries (e.g., EU member states (Asmild and Zhu, 2016), Spain (Lozano-Vivas, 1997), China (Asmild and Matthews, 2012) and South Asia (Bibi *et al.*, 2018)), there is a lot of potential for further banking applications of our approach, which we provide some practical guidance on.

The remainder of this paper is organized as follows. Section 2 has two parts. In the first part we set out the general form of our spatial cost model and provide a brief overview of our model estimation strategy. In the second part we explain the method that underlies the direct, composite indirect and total elasticities, and we also set out the three-part decomposition of a composite indirect elasticity. Based on these elasticities, in section 3 we set out our method to calculate the spatial RSE, $EPSE$ and $EPSU$ measures. In section 4 we apply these new measures to U.S. banks for the period $1998 - 2015$ and discuss the policy implications. Section 5 concludes and elaborates on the scope for further applications.

2 Cost Function with Simultaneous Spillover Regimes

2.1 Model Layout and an Overview of the Estimation Procedure

We estimate a panel data spatial Durbin cost function (SDCF) with fixed effects and simultaneous spillover regimes. Our structural model has the following general form, where the variables are logged.

$$c_{it} = \alpha + TL\left(y_{it}, p_{it}, t_i\right) + \left(\begin{array}{c} STL_1\left(\sum\limits_{j=1}^{N} w_{ij1}y_{jt}, \sum\limits_{j=1}^{N} w_{ij1}p_{jt}, \sum\limits_{j=1}^{N} w_{ij1}t_j\right) + ... \\ +STL_M\left(\sum\limits_{j=1}^{N} w_{ijM}y_{jt}, \sum\limits_{j=1}^{N} w_{ijM}p_{jt}, \sum\limits_{j=1}^{N} w_{ijM}t_j\right) \end{array} \right) + \left(\begin{array}{c} \delta_1 \sum\limits_{j=1}^{N} w_{ij1}c_{jt} + ... \\ +\delta_M \sum\limits_{j=1}^{N} w_{ijM}c_{jt} \end{array} \right) + \kappa_i + \varepsilon_{it}. \tag{1}$$

The panel data comprises observations for N firms and T periods, which are indexed $i, j = 1, ..., N$ $\forall$ i, j and $t = 1, ..., T$. Following the typical case that is encountered when using firm level data we take N to be large and T to be small. c_{it} is an observation for total cost for the ith firm in period t and together with $TL\left(y_{it}, p_{it}, t_i\right) = \rho t_i + \frac{1}{2}\varsigma t_i^2 + \zeta' p_{it} + \xi' y_{it} + \frac{1}{2}p'_{it}\mathbf{\Theta}p_{it} + \frac{1}{2}y'_{it}\mathbf{\Gamma}y_{it} + p'_{it}\mathbf{\Psi}y_{it}$ represents the variable returns to scale translog approximation of the log of the cost frontier technology. α is the common intercept; y_{it} is the $(1 \times K)$ vector of observations for the outputs, which are indexed $k = 1, ..., K$; p_{it} is the $(1 \times L)$ vector of observations for the input prices, which are indexed $l = 1, ..., L$;

t and t^2 collectively represent a non-linear time trend; ε_{it} is the idiosyncratic error; and κ_i is a fixed effect.[5] We account for unobserved heterogeneity using fixed effects rather than random effects because the former are less restrictive as they can be correlated with the time-varying errors.

In Eq. 1 there are M simultaneous spillover regimes. M corresponds to the number of spatial weights matrices indexed $\mathbf{W_m}= \mathbf{W_1}, ..., \mathbf{W_M}$, where $\mathbf{W_m}$ is an $(N \times N)$ exogenous matrix that is specified in advance of estimating the model. $\mathbf{W_m}$ comprises non-negative constant $i, j-$th elements w_{ijm}, which are commonly referred to as spatial weights. In practice, M will be small reflecting a small number of very distinct sets of spatial linkages, as opposed to quite a large number of rather similar competing spatial weights matrices, where there is a high chance of collinearity between the resulting spatial autoregressive (SAR) variables. Accordingly, there are two simultaneous spatial regimes in our application to U.S. banks, i.e., the spatial linkages between banks' full service brick and mortar branch networks and the spatial linkages between the networks of all their other branch types collectively.

In general terms, we model the cross-sectional correlation using multiple simultaneous spatial processes that explicitly relate a firm to its neighbors (i.e., $\mathbf{W_1}, ..., \mathbf{W_M}$), which for the simpler case of a single spatial process is an approach with a long history (see Whittle, 1954). More specifically, for the mth spatial regime $\mathbf{W_m}$ represents the spatial arrangement of the firms and the strength of the interaction between the firms. As a result of the latter, all the elements on the main diagonal of $\mathbf{W_m}$ are set to zero as a firm cannot be part of its own neighborhood set. A measure of proximity must be used to populate $\mathbf{W_m}$, where in line with the situation for nearly all spatial models we are conscious that our model takes $\mathbf{W_1}, ..., \mathbf{W_M}$ to be exogenous. With this in mind we use what can be taken in the context of our application to be a exogenous measure of geographical proximity. As will become apparent though the measure of geographical proximity we use is rather novel.

Although there are clearly geographical links between banks that operate in the same markets because they compete with one another and are exposed to the same headline changes in local and regional economies, there are of course also economic linkages between banks. One would suspect that most measures of economic proximity between banks such as inter-bank loans would be endogenous, but it may be possible to come up with an exogenous measure. A test has recently been developed for an endogenous spatial weights matrix (Bera *et al.*, 2018), but it is not clear how we would proceed with our analysis using endogenous specifications of $\mathbf{W_1}, ..., \mathbf{W_M}$. This is because despite some progress by Qu and Lee (2015) on accounting for an endogenous spatial weights matrix by using the control function method to propose three new estimators (a two-stage instrumental variables (2SIV) method, a generalized method of moments (GMM) approach and a quasi-maximum likelihood (QML) procedure), these estimators are for cross-sectional data only. To incorporate into our banking modeling framework an endogenous spatial weights matrix, one would need to extend their GMM/QML estimator to panel data with fixed effects, which is outside the scope of this paper.[6] We are therefore prudent and ensure that we avoid the situation where $\mathbf{W_1}, ..., \mathbf{W_M}$ are endogenous by specifying these matrices using an exogenous measure of geographical proximity.

Since our model is based on the literature that revolves around the spatial weights matrix, which

[5] As we model multiple spatial regimes for parsimony we omit from our specification of $TL\left(y_{it}, p_{it}, t_i\right)$ interactions with t.

[6] More generally, further work on spatial econometric models with an endogenous spatial weights matrix based on economic/social linkages would draw parallels with the modeling of these linkages in other types of econometric analysis. These include time series studies that look at the impact of a well-established economic linkage (e.g., Pesaran *et al.*, 2004), and studies that consider a range of social linkages to investigate which are the sources of the spatial correlation (e.g., Conley and Topa, 2002, which is based in part on the methodology in Conley, 1999). Our paper is more akin to the approach in the time series literature as we use indirect elasticities to quantify the impact of the spatial correlation from the geographical linkage between banks that have overlapping branch networks.

is one of the two main strands of the literature on panel data modeling of cross-sectional dependence when N is large relative to T, we account for what is referred to as the weak form of spatial dependence. This is because the multiple simultaneous spatial processes of the dependent variable are characterized by a fading memory across space. The fading memory arises because underlying our model is the assumption that the row and column sums of $\mathbf{W_1}, ..., \mathbf{W_M}$ before normalization (denoted as $\widetilde{\mathbf{W}}_\mathbf{1}, ..., \widetilde{\mathbf{W}}_\mathbf{M}$) and the row and column sums of $(\mathbf{I} - \delta_1 \, \widetilde{\mathbf{W}}_\mathbf{1} - ... - \delta_M \widetilde{\mathbf{W}}_\mathbf{M})^{-1}$ (for all values of $\delta_1, ..., \delta_M$) are uniformly bounded in absolute value as $N \to \infty$. As a result of this fading memory assumption, the spatial processes are limited to a manageable degree which rules out explosive growth of the dependent variable across space (Kelejian and Prucha, 2001).

The other main strand of the above literature accounts for the strong form of spatial dependence, which differs from its weak counterpart as there is no decay in the dependence across space. This literature uses multi-factor models comprising a number of unobserved common factors that lead to multiple common effects of different magnitudes on all the spatial units. These models are well-suited to analyze the common effects across U.S. banks of industry wide phenomena such as regulatory policies and the financial crisis. Although analyzing the common effects across space of such macro phenomena is clearly interesting, our focus is on the external economies of scale that gravitate to a bank from other banks, which is an interesting different type of phenomenon. It is different because these external economies consider bank interdependence at the micro level which, as a result, varies across space according to the extent to which banks operate in the same markets. A multi-factor model would not take this variation in the dependence across space in to account, so we instead use a model that is based on the spatial weights matrix and which also yields the indirect elasticities that our new external economies are based on.

Having specified $\mathbf{W_m}$ one can construct spatial lags of the dependent and independent variables. The mth spatial lag of the dependent variable, $\sum_{j=1}^{N} w_{ijm} c_{jt}$, is endogenous, which our estimator accounts for. We adopt a sufficiently general parameter space for the associated SAR parameter, $\delta_m \in \left[\frac{1}{\min(r_1^{\min},...,r_M^{\min})}, \frac{1}{\max(r_1^{\max},...,r_M^{\max})}\right]$, where $r_m^{\max}$ is the most positive real characteristic root of $\mathbf{W_m}$. Note that $\mathbf{W_m}$ denotes a normalized specification of our mth spatial weights matrix, where the normalization we use in our empirical application gives $r_1^{\max}, ..., r_M^{\max} = 1$. For details of this normalization see subsection 4.1. In our empirical application $\mathbf{W_m}$ before normalization is asymmetric so $\mathbf{W_m}$ may have complex roots. In this case LSP prove that the lower limit of δ_m is the inverse of $r_m^{\min}$, which is the most negative *purely* real characteristic root of $\mathbf{W_m}$. Following the spatial literature (e.g., Anselin, 2003), Eq. 1 is the structural form of our model as it includes the M SAR variables as regressors. In the reduced form of the model, which we use to compute the elasticities, these SAR variables do not feature.

In Eq. 1 $STL_1 + ...+ STL_M$ represent M spatial lags of $TL\,(y_{it}, p_{it}, t_i)$ and together with $\sum_{j=1}^{N} w_{ijm} c_{jt}$ they shift the frontier technology.[7] In contrast to $\sum_{j=1}^{N} w_{ijm} c_{jt}$, which accounts for endogenous global SAR dependence, STL_m is exogenous in Eq. 1 and accounts for local spatial dependence. If we were to omit $STL_1 + ...+ STL_M$ from Eq. 1 the model collapses to what is referred to as the SAR model with multiple spatial regimes. We include $STL_1 + ...+ STL_M$ though because in the empirical spatial literature local spatial variables are frequently found to be important determinants. As a result of the local spatial variables in our model it is referred to as the spatial Durbin model with multiple spatial regimes.

[7] If we were to use a row-normalized specification of $\mathbf{W_m}$ we would omit from the mth spatial lag of $TL\,(y_{it}, p_{it}, t_i)$, which we denote STL_m, $\sum_{j=1}^{N} w_{ijm} t_j$ and $\sum_{j=1}^{N} w_{ijm} t_j^2$ due to perfect collinearity with t_i and t_i^2 (e.g., $t_i = \sum_{j=1}^{N} w_{ijm} t_j$). In our application we include $\sum_{j=1}^{N} w_{ijm} t_j$ and $\sum_{j=1}^{N} w_{ijm} t_j^2$ because the $\mathbf{W_m}$ we use, as will become apparent in subsection 4.1, involves an alternative normalization and, as a result, there is no such collinearity.

In addition to the aforementioned parameters there are additional parameters (ρ, $\frac{1}{2}\varsigma$, ρ_{sm} and $\frac{1}{2}\varsigma_{sm}$, where a subscript s denotes a local spatial parameter), vectors of parameters (ζ', ξ', ξ'_{sm} and ζ'_{sm}) and matrices of parameters ($\frac{1}{2}\mathbf{\Theta}$, $\frac{1}{2}\mathbf{\Gamma}$, $\mathbf{\Psi}$, $\frac{1}{2}\mathbf{\Theta_{sm}}$, $\frac{1}{2}\mathbf{\Gamma_{sm}}$ and $\mathbf{\Psi_{sm}}$) to be estimated. From the properties of the translog functional form (Christensen *et al.*, 1973) Eq. 1 is twice differentiable with respect to an output, an input price and their M spatial lags. Due to the symmetry restrictions that are placed on the parameter matrices the resulting Hessians are also symmetric.

Parametric estimation of a spatial models involves using ML, 2SIV, GMM or Bayesian Monte Carlo Markov Chain methods. We follow, for example, Elhorst and Fréret (2009) and Elhorst *et al.* (2012) and estimate our model using ML. The estimation of our model has a couple of important features. First, as is standard for fixed effects models, we estimate our model using the within transformation by demeaning the variables at the level of each firm which eliminates the fixed effects. This transformation circumvents the well-known incidental parameter problem associated with the fixed effects. Second, the log-likelihood function includes the scaled logged determinant of the Jacobian of the transformation from ε^*_{it} to c^*_{it}, where the $*$ denotes demeaned transformations of ε_{it} and c_{it}. In other words, the log-likelihood function includes $T \log |\mathbf{I} - \delta_1 \mathbf{W_1} - ... - \delta_m \mathbf{W_m}|$. As is standard in spatial econometrics and mirroring the role of the transformation from ε^*_{it} to c^*_{it} in ML estimation of the simpler model with a single SAR variable, the transformation from ε^*_{it} to c^*_{it} for our model accounts for the endogeneity of the SAR variables and the fact that ε_{it} is not observed (Anselin, 1988; Elhorst, 2009).

2.2 Method for the Elasticities and the Decomposition

Due to LSP it is now well-known for the structural form of a model in logs with at least one SAR variable such as Eq. 1 that the fitted parameters cannot be interpreted as elasticities. The reason is that the elasticity of a variable is a function of the SAR parameter(s). To compute the elasticities for the variables for such a model we adopt what is now the standard approach in spatial econometrics, which involves computing the direct, indirect and total elasticities using the fitted parameters from the structural form of the model. A direct elasticity is interpreted in the same way as an elasticity from a non-spatial model, although a direct elasticity takes into account feedback effects that occur via the spatial multiplier matrix. Feedback is the effect of a change in an independent variable for a particular firm which reverberates back to the same firm's dependent variable through its effect on the dependent variables of the other firms in the sample. An indirect elasticity can be calculated in two ways yielding the same numerical value. This leads to two interpretations of an indirect elasticity: (i) average change in the dependent variable of all the other firms in the sample following a change in an independent variable for one firm; or (ii) average change in the dependent variable for one firm following a change in an independent variable for all the other firms in the sample. Summing the direct and indirect elasticities gives the total elasticity.

The indirect elasticity we refer to above is a composite measure. We now explain how we compute the direct, composite indirect and total elasticities and their $t-$statistics. As spillovers is the focus of our interest we also explain how we decompose a composite indirect elasticity into its constituent parts. Direct, composite indirect and total elasticities are computed from the reduced form of a structural spatial model. We obtain the reduced form of Eq. 1 by first rewriting the model using matrix notation as follows.

$$c_t = \alpha\iota + TL\left(y_t, p_t, t\right) + \left(\begin{array}{c} STL_1\left(\mathbf{W_1}y_t, \mathbf{W_1}p_t, \mathbf{W_1}t\right) + ... \\ +STL_M\left(\mathbf{W_M}y_t, \mathbf{W_M}p_t, \mathbf{W_M}t\right)\end{array}\right) + \left(\begin{array}{c} \delta_1\mathbf{W_1}c_t + ... \\ +\delta_M\mathbf{W_M}c_t\end{array}\right) + \kappa + \varepsilon_t, \tag{2}$$

where the i and j subscripts from Eq. 1 are dropped to denote vectors of successively stacked cross-sectional observations, ι is an $(N \times 1)$ vector of ones and everything else is as previously defined for Eq. 1. We next take $(\delta_1\mathbf{W_1}c_t + ... + \delta_M\mathbf{W_M}c_t)$ to the left-hand side which gives $(\mathbf{I} - \delta_1\mathbf{W_1} - ... - \delta_M\mathbf{W_M})\,c_t$, and then take $(\mathbf{I} - \delta_1\mathbf{W_1} - ... - \delta_M\mathbf{W_M})$ to the right-hand side to obtain the reduced form of the model in Eq. 3.

$$c_t = \left(\begin{array}{c} \mathbf{I} - \delta_1\mathbf{W_1} - ... \\ -\delta_M\mathbf{W_M}\end{array}\right)^{-1}\left(\begin{array}{c} \alpha\iota + TL\left(y_t, p_t, t\right) + \left(\begin{array}{c} STL_1\left(\mathbf{W_1}y_t, \mathbf{W_1}p_t, \mathbf{W_1}t\right) + ... \\ +STL_M\left(\mathbf{W_M}y_t, \mathbf{W_M}p_t, \mathbf{W_M}t\right)\end{array}\right) \\ +\kappa + \varepsilon_t\end{array}\right), \tag{3}$$

where using terminology from the spatial literature $(\mathbf{I} - \delta_1\mathbf{W_1} - ... - \delta_M\mathbf{W_M})^{-1}$ is the spatial multiplier matrix, which is key in spatial models.

The intuition behind the above mathematical transition from the structural form of the model to its reduced form is as follows. In the structural form the spatial correlations between firms' costs across the M networks is accounted for by the M SAR variables. Additionally, in the structural form the spatial lags of a firm's own independent variables capture the spatial correlations between the effects of a firm's own independent variables on its costs across neighboring firms in the M networks. By transforming the structural form into its reduced form we are simply attributing these different spatial correlations to the own independent variables to obtain a form of the model that yields interpretable elasticities.

We present the method to calculate the direct, composite indirect and total elasticities for the variables in the context of a first order output, which we denote $y_{k,t}$. To facilitate the presentation of this method we distinguish between the structural form of our model in Eq. 1 and its local spatial counterpart, i.e., Eq. 1 with the M SAR variables omitted. The local spatial counterpart would only account for first order neighbor effects via the exogenous spatially lagged variables, whereas Eq. 1 is a global spatial model due to the presence of the SAR variable(s). If we were to estimate the local spatial counterpart using mean adjusted data, which is a common data transformation for translog functions, all the fitted coefficients on the first order own and local spatial variables are elasticities at the sample mean. This is because at the sample mean the own and local spatial quadratic and interaction terms are zero. Applying this to the reduced form of our model in Eq. 3, the ξ_k and $\xi_{s1,k}, ..., \xi_{sM,k}$ coefficients on the $y_{k,t}$ and $\mathbf{W_1}y_{k,t}, ..., \mathbf{W_M}y_{k,t}$ variables can be directly used to calculate the direct, composite indirect and total elasticities for $y_{k,t}$ at the sample mean. Note that the first order direct, composite indirect and total parameters are only elasticities at the sample mean and not at any other points in the sample. To obtain elasticities outside the sample mean we must apply the approach for a non-spatial translog function to our spatial setting. This involves using, among other things, the direct, composite indirect and total parameters on the first order, quadratic and interaction variables.

If, as set out in Eq. 4a, we differentiate Eq. 3 with respect to $y_{k,t}$, we obtain a matrix of direct and composite indirect elasticities for individual firms from the product of the two matrices on the right-hand side of Eq. 4b, where this product is independent of the time index. To summarize the large number of elasticities from this product we report the mean direct elasticity (mean of the diagonal elements of this product) and a mean composite indirect spillover elasticity. (The mean spillover elasticity *to* a firm is the mean row sum of the off-diagonal elements of this product and the mean spillover elasticity *from* a firm is the mean column sum of the off-diagonal elements). Across all the firms in the sample it does not matter which of the two measures of the mean composite indirect elasticity we use as they yield the same numerical value. In contrast, outside the sample mean for a subset of the firms (e.g., the mean for a particular quartile), the two measures of the mean composite indirect elasticity will not be equal. In this case we use the former measure of the mean composite indirect elasticity as we are interested in the spillover of scale and product mix economies that gravitate to a firm.

$$\left[\begin{array}{ccc} \frac{\partial c}{\partial y_{k,1}}, & \cdots & , \frac{\partial c}{\partial y_{k,N}} \end{array}\right]_t = \left[\begin{array}{ccc} \frac{\partial c_1}{\partial y_{k,1}} & \cdots & \frac{\partial c_1}{\partial y_{k,N}} \\ \vdots & \ddots & \vdots \\ \frac{\partial c_N}{\partial y_{k,1}} & \cdots & \frac{\partial c_N}{\partial y_{k,N}} \end{array}\right]_t \tag{4a}$$

$$= \left(\mathbf{I} - \delta_1 \mathbf{W_1} - ... - \delta_M \mathbf{W_M}\right)^{-1} \times \left[\begin{array}{ccc} \xi_k & \cdots & \begin{array}{c} w_{1N,1}\xi_{s1,k} + ... \\ +w_{1N,M}\xi_{sM,k} \end{array} \\ \vdots & \ddots & \vdots \\ \begin{array}{c} w_{N1,1}\xi_{s1,k} + ... \\ +w_{N1,M}\xi_{sM,k} \end{array} & \cdots & \xi_k \end{array}\right] \tag{4b}$$

$$= \left(\mathbf{I} - \delta_1 \mathbf{W_1} - ... - \delta_M \mathbf{W_M}\right)^{-1} \times \left(\begin{array}{c} \left[\begin{array}{ccc} \xi_k & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \xi_k \end{array}\right] + \left[\begin{array}{ccc} 0 & \cdots & w_{1N,1}\xi_{s1,k} \\ \vdots & \ddots & \vdots \\ w_{N1,1}\xi_{s1,k} & \cdots & 0 \end{array}\right] + \\ ... + \left[\begin{array}{ccc} 0 & \cdots & w_{1N,M}\xi_{sM,k} \\ \vdots & \ddots & \vdots \\ w_{N1,M}\xi_{sM,k} & \cdots & 0 \end{array}\right] \end{array}\right). \tag{4c}$$

In the context of the $y_{k,t}$ variable we are using for illustrative purposes, we can see from the reduced form of the model (Eq. 3) that this variable is multiplied by the direct, composite indirect and total elasticities. A direct elasticity is akin to an own elasticity from a standard non-spatial model because the only difference between the two is the feedback component in the direct elasticity, which originates from the *ith* firm anyway and is typically small in empirical applications. We therefore refer to the direct elasticity multiplied by $y_{k,t}$ as a change in the firm's output that is under its control (i.e., a change in a firm's output that is considered in textbook production theory). In contrast, we refer to the composite indirect elasticity multiplied by $y_{k,t}$ as a change in the firm's output due to the spillover effect from other firms, which is primarily, but not entirely, outside the firm's control (i.e., a change in a firm's output that is not considered in textbook production theory).

We follow the spatial literature by calculating $t-$statistics for the mean direct, composite indirect and total parameters via Monte Carlo simulation of the distributions of these means. This involves drawing $1,000$ parameter combinations from the variance-covariance matrix for Eq. 1, where each parameter has a random component drawn from $N\left(0,1\right)$. For each parameter combination we calculate mean direct, composite indirect and total parameters. The $t-$statistics for the mean direct, composite indirect and total parameters from Eq. 4b are then calculated by dividing each of these mean parameters by the standard error of the parameter across the $1,000$ estimates from the Monte Carlo simulations.[8]

We use Eq. 4c to decompose a composite indirect elasticity into $M+1$ components. This involves using the $M+1$ constituent parts of the second matrix on the right-hand side of Eq. 4b. The $M+1$ decomposed indirect elasticities comprise the following. (i) An indirect elasticity that measures the spillover to a firm's dependent variable due to the spillover effect that is attributed to its independent variable from all M spatial lags of the dependent variable, where these lags are accounted for in the spatial multiplier matrix in Eq. 4c by the matrices $-\delta_1\mathbf{W_1} - ... - \delta_M\mathbf{W_M}$. (ii) M indirect spillover elasticities that measure the spillovers to a firm's dependent variable due to the M spillover effects that are attributed to its independent variable from each of the M spatial lags of the latter.

(i) relates to the off-diagonal elements of the product of the spatial multiplier matrix and the matrix with ξ_k along the main diagonal in Eq. 4c. In terms of our application, (i) measures the cost spillover to a bank due to the spillover effect that is attributed to its independent variable from both spatial lags of the dependent cost variable, which together capture the spatial correlations of banks' costs across the banks' networks of brick and mortar branches and other types of branches.

The *mth* indirect elasticity from (ii) relates to the off-diagonal elements of the product of the spatial multiplier matrix and the *mth* matrix in Eq. 4c with $w_{ij,m}\xi_{sm,k}$ off-diagonal elements. The two measures of (ii) that we calculate in our application measure the cost spillover to a bank due to the spillover effect that is attributed to its independent variable from the each of the two spatial lags of the latter. These two spatial lags and thus the two cost spillover elasticities capture the spatial correlations between the effects of a bank's independent variable on its cost across the banks' networks of brick and mortar branches and other types of branches. As described above for mean direct, composite indirect and total parameters, the $t-$statistics for the three types of mean decomposed indirect parameters we report (i.e., from above one measure of (i) and two measures of (ii)) are calculated via Monte Carlo simulation.

3 Spatial Measures of Scale and Product Mix Economies

From the direct, composite indirect and total translog functions in Eqs. $5-7$, which we construct using the relevant parameters from Eq. 4b, we compute the spatial measures of returns to scale and product mix.

$$c_{it}^{Dir} = \rho^{Dir}t_i + \frac{1}{2}\varsigma^{Dir}t_i^2 + \zeta^{Dir'}p_{it} + \xi^{Dir'}y_{it} + \frac{1}{2}p'_{it}\boldsymbol{\Theta}^{Dir}p_{it} + \frac{1}{2}y'_{it}\boldsymbol{\Gamma}^{Dir}y_{it} + p'_{it}\boldsymbol{\Psi}^{Dir}y_{it}, \tag{5}$$

$$c_{it}^{CInd} = \rho^{CInd}t_i + \frac{1}{2}\varsigma^{CInd}t_i^2 + \zeta^{CInd'}p_{it} + \xi^{CInd'}y_{it} + \frac{1}{2}p'_{jt}\boldsymbol{\Theta}^{CInd}p_{it} + \frac{1}{2}y'_{it}\boldsymbol{\Gamma}^{CInd}y_{it} + p'_{jt}\boldsymbol{\Psi}^{CInd}y_{it}, \tag{6}$$

$$c_{it}^{Tot} = \rho^{Tot}t_i + \frac{1}{2}\varsigma^{Tot}t_i^2 + \zeta^{Tot'}p_{it} + \xi^{Tot'}y_{it} + \frac{1}{2}p'_{it}\boldsymbol{\Theta}^{Tot}p_{it} + \frac{1}{2}y'_{it}\boldsymbol{\Gamma}^{Tot}y_{it} + p'_{it}\boldsymbol{\Psi}^{Tot}y_{it}, \tag{7}$$

[8] We thank an anonymous reviewer for suggesting that we elaborate on how we calculate the $t-$statistics for the mean direct, composite indirect and total parameters.

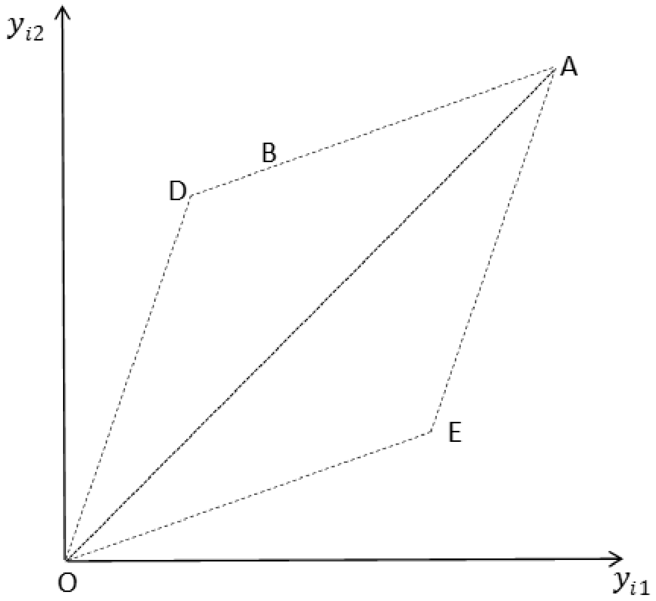
where Dir, $CInd$ and Tot denote direct, composite indirect and total. In other words, we compute direct, composite indirect and total returns to scale and product mix. In the same way we also construct $M+1$ decomposed indirect translog functions, which we use to calculate decomposed indirect measures of returns to scale and product mix. The discussion of the indirect translog functions, however, is confined to the composite case as it is simple to adapt the presentation to the $M+1$ decomposed indirect translog functions. This involves replacing the composite indirect parameters in Eq. 6 with each set of decomposed indirect parameters.

It is evident from Eq. 6 that a composite indirect elasticity measures the effect on c_{it}^{CInd} from the spillover impact that permeates to the ith firm's independent variable. Although Eqs. $5-7$ all have a translog functional form and from the properties of this functional form the Hessians from these equations are symmetric, the theoretical monotonicity and curvature properties only apply to the direct translog function and not to the composite and decomposed indirect and total functions. This is because a direct parameter is akin to an own parameter from a standard non-spatial model, whereas textbook production theory does not say anything about composite/decomposed spillovers and hence the total parameters. In our application we check the proportion of the sets of direct output and input price elasticities over the sample that satisfy the monotonicity property, and the proportion of the sample where our direct translog cost function satisfies the curvature property. In addition to the mismatch between the theoretical properties of a non-spatial translog function vis-à-vis composite/decomposed indirect and total functions, unlike for a non-spatial cost function where of course we observe the dependent variable, we do not observe c_{it}^{Dir}, c_{it}^{CInd} and c_{it}^{Tot}. Using Eqs. $5-7$ though one can compute c_{it}^{Dir}, c_{it}^{CInd} and c_{it}^{Tot}. Moreover, in contrast to, for example, the M SAR variables in Eq. 1, w_{ij} does not pre-multiply the observations in Eqs. $5-7$ because in these equations the effect of the spatial weights is incorporated within the estimates of the direct, composite indirect and total parameters.

To set the scene for the spatial scale and product mix economies consider figure 1, which is a slight modification of a figure in Berger *et al.* (1987) and Wheelock and Wilson (2001) for the non-spatial case. In the ith firm's two-dimensional output space and using our terminology, their figure is in terms of the bundle of output levels that are under the firm's control (y_{i1} and y_{i2}). In contrast, our figure is for the composite indirect case as the source of changes in y_{i1} and y_{i2} in figure 1 is composite spillovers to the ith firm, which are primarily, but not entirely, outside the firm's control. We can also apply figure 1 to: (i) the direct case if the source of changes in y_{i1} and y_{i2} is the ith firm; (ii) a decomposed indirect case if the source of changes in y_{i1} and y_{i2} is decomposed spillovers to the ith firm; and (iii) the total case for changes in y_{i1} and y_{i2} that are from the combined sources of the ith firm and composite spillovers to the firm.

We extend the non-spatial ray, expansion-path and expansion-path subadditivity measures of scale and product mix economies in banking (Berger *et al.*, 1987; Wheelock and Wilson, 2001) to the spatial case. More specifically, we extend these non-spatial measures for banks to the case where there are simultaneous spillover regimes. This involves introducing new sets of measures of scale and product mix economies, where each set comprises direct, composite and decomposed indirect and total measures. We set out formally below each new set of measures of the scale and product mix economies in the context of the composite indirect measure as it is simple to adapt the presentation to obtain the direct, decomposed indirect and total measures.

Using figure 1 we illustrate output levels for the three types of composite indirect economies as follows, where as we noted above the sources of the changes in y_{i1} and y_{i2} are composite spillovers to the ith firm. (a) Composite indirect ray-scale economies (RSE^{CInd}) relate to equiproportional



Note this figure assumes that the source of changes in the output levels is the composite spillovers to the *ith* firm.

Figure 1: Illustrative output levels for the composite indirect spillover scale and product mix economies

changes in y_{i1} and y_{i2} along the radial ray **OA**. (b) Composite indirect expansion-path scale economies $(EPSE^{CInd})$ relate to incremental changes in y_{i1} and y_{i2} along a non-radial ray such as **DA**, which represents a portion of the firm's composite indirect output expansion-path **ODA**. (c) Composite indirect expansion-path subadditivity $(EPSU^{CInd})$ measures the returns from a single firm at **A** producing the combined outputs of two smaller firms at **D** and **E**.

To calculate the direct, composite and decomposed indirect and total RSE, each radial ray is characterized by the firm producing the relevant type of outputs in constant proportion to one another, as defined by the fitted model. In other words, the product mix of the *ith* firm's relevant type of outputs is held constant along the ray. Whereas with these RSE measures the absolute mix of the relevant type of outputs does not change along the ray, when we calculate the direct, composite and decomposed indirect and total $EPSE$ this mix changes along the relevant non-radial output expansion path. There will though be no change along an expansion path in the relative proportions of the relevant type of outputs.

With reference to figure 1 we now turn to the formal presentation of the composite indirect measures of the scale and product mix economies.[9]

1. *Spatial Ray-Scale Economies (RSE)*

Consider a particular point $\left(y_i^{\mathbf{O}}, p_i^{\mathbf{O}}, t_i^{\mathbf{O}}\right)$ in the (y_i, p_i, t_i) space, where the set of points $\Re = \left[\left(\phi y_i^{\mathbf{O}}, p_i^{\mathbf{O}}, t_i^{\mathbf{O}}\right) | \phi \in (0, \infty)\right]$ represents a radial ray. The superscript **O** in the context of the two-dimensional output space in figure 1 represents the output levels at a point along **OA**.

From a theoretical perspective, to measure RSE^{CInd} along $\Re$ we define $\mathcal{F}_i^{CInd}(\phi|y_i) \equiv c_i^{CInd}(\phi y_i, p_i, t_i) / \phi c_i^{CInd}(y_i, p_i, t_i)$. If $\mathcal{F}_i^{CInd}(\phi|y_i)$ is increasing (/constant/decreasing) in ϕ, RSE^{CInd}

[9] We thank an anonymous reviewer for suggesting that we relate our formal presentation of the composite indirect scale and product mix economies to figure 1.

is increasing (/constant/decreasing) along $\Re$. By adapting the approach in the large body of empirical literature that estimates non-spatial RSE, for a particular point $\left(y_i^{\mathbf{O}}, p_i^{\mathbf{O}}, t_i^{\mathbf{O}}\right)$ along $\Re$ we compute RSE_i^{CInd} using Eq. 8.

$$RSE_i^{CInd} = \frac{\partial c_i^{CInd}\left(\phi y_i, p_i, t_i\right)}{\partial \phi}|_{\phi=1} = \sum_{k=1}^{K} \frac{\partial c_i^{CInd}\left(y_i, p_i, t_i\right)}{\partial y_{ik}}, \tag{8}$$

where the elasticity $\partial c_i^{CInd}\left(y_i, p_i, t_i\right)/\partial y_{ik}$ is the first order derivative of Eq. 6 with respect to the kth output of the ith firm.

In order to consider in the application RSE_i^{Dir}, RSE_i^{CInd} and RSE_i^{Tot} at different points along the relevant radial ray, we compute these RSE measures at the sample mean and at various other points in the sample. The resulting estimates indicate how expected c_i^{Dir}, c_i^{CInd} and c_i^{Tot} vary along the relevant radial ray, which informs how RSE_i^{Dir}, RSE_i^{CInd} and RSE_i^{Tot} vary at different scales of production.

For the direct, composite indirect and total returns in each of the three sets of spatial scale and product mix economies, the classification of the returns is the same as in the standard non-spatial case. Accordingly, RSE_i^{Dir}, RSE_i^{CInd} and RSE_i^{Tot} $<$, $=$ or $>$ 1 indicates increasing, constant or decreasing direct, composite indirect and total returns to scale. The classification of RSE_i^{Dir}, RSE_i^{CInd} and RSE_i^{Tot} and also the direct, composite indirect and total returns from the other two sets of spatial scale and product mix economies need not of course be the same. We conduct statistical inference by using the direct, composite indirect and total parameters from the Monte Carlo simulations to compute $1,000$ estimates of RSE_i^{Dir}, RSE_i^{CInd} and RSE_i^{Tot}.

Summarizing: if RSE_i^{Dir}, $\left(RSE_i^{CInd}\right)$, $\left[RSE_i^{Tot}\right]$ is $<$, $=$ or $>$ 1 and there is an equiproportional increase in each of the ith firm's K outputs along the relevant radial ray, where the source of each output is the ith firm (composite spillovers to the ith firm) [the ith firm and composite spillovers to the firm combined], c_i^{Dir}, $\left(c_i^{CInd}\right)$, $\left[c_i^{Tot}\right]$ will rise by a smaller, the same or a larger proportion.

2. *Spatial Expansion-Path Scale Economies (EPSE)*

Although non-spatial and spatial RSE are convenient metrics and in the non-spatial setting is a widely reported measure, in practice these metrics may not be appropriate as a firm is unlikely to be located along the relevant radial ray. To address this shortcoming Berger *et al.* (1987) propose a non-spatial measure of the degree of scale economies along a firm's output expansion-path ($EPSE$).

To move from non-spatial $EPSE$ to the spatial case of $EPSE^{CInd}$ we consider another point $\left(y_i^{\mathbf{B}}, p_i^{\mathbf{B}}, t_i^{\mathbf{B}}\right)$ in the (y_i, p_i, t_i) space, where $\left(y_i^{\mathbf{B}}, p_i^{\mathbf{B}}, t_i^{\mathbf{B}}\right)$ lies somewhere along a non-radial ray. $EPSE^{CInd}$ measures the increase in expected c_i^{CInd} as the firm moves along this ray between the points $\left((1-\lambda)\, y_i^{\mathbf{B}}, p_i^{\mathbf{B}}, t_i^{\mathbf{B}}\right)$ and $\left((1+\lambda)\, y_i^{\mathbf{B}}, p_i^{\mathbf{B}}, t_i^{\mathbf{B}}\right)$, where λ is a small positive number that we elaborate on below. A movement in figure 1 that is representative of the move between the points $\left((1-\lambda)\, y_i^{\mathbf{B}}, p_i^{\mathbf{B}}, t_i^{\mathbf{B}}\right)$ and $\left((1+\lambda)\, y_i^{\mathbf{B}}, p_i^{\mathbf{B}}, t_i^{\mathbf{B}}\right)$ would be a move between points that represent incremental moves from point **B** down and up the non-radial ray **DA**.

Using c_i^{CInd} from Eq. 6, we compute $EPSE_i^{CInd}$ as follows.

$$EPSE_i^{CInd} = \frac{c_i^{CInd}\left(\phi\,(1-\lambda)\, y_i^{\mathbf{B}}, p_i^{\mathbf{B}}, t_i^{\mathbf{B}}\right)}{\phi c_i^{CInd}\left((1-\lambda)\, y_i^{\mathbf{B}}, p_i^{\mathbf{B}}, t_i^{\mathbf{B}}\right)} = \frac{c_i^{CInd}\left((1+\lambda)\, y_i^{\mathbf{B}}, p_i^{\mathbf{B}}, t_i^{\mathbf{B}}\right)}{\left(\frac{1+\lambda}{1-\lambda}\right) c_i^{CInd}\left((1-\lambda)\, y_i^{\mathbf{B}}, p_i^{\mathbf{B}}, t_i^{\mathbf{B}}\right)}, \tag{9}$$

where due to the relative proportions of the firm's composite indirect spillovers of outputs being constant $\phi\,(1-\lambda)\, y_i^{\mathbf{B}} = (1+\lambda)\, y_i^{\mathbf{B}}$. This gives $\phi = (1+\lambda)/(1-\lambda)$ and subsequently the expression on the far right of Eq. 9. As the equations for $EPSE_i^{Dir}$, $EPSE_i^{CInd}$ and $EPSE_i^{Tot}$ have the same form it is simple to adapt Eq. 9 to the cases of $EPSE_i^{Dir}$ and $EPSE_i^{Tot}$.

Following the non-spatial study by Wheelock and Wilson (2012), in the empirical application we use $\lambda = 0.05$ to compute $EPSE_i^{Dir}$, $EPSE_i^{CInd}$ and $EPSE_i^{Tot}$. In particular, we compute these $EPSE$ measures for movements along the relevant output expansion-path between $\pm\lambda$ of the mean output vector for the full sample or a sub-sample. We therefore consider movements between 95% and 105% of the relevant mean output vector.

Clearly if each of the *ith* firm's K outputs increases by a factor $\phi > 1$, where the source of the increase is the *ith* firm, composite spillovers to the firm and the combined sources of the *ith* firm and composite spillovers to the firm, c_i^{Dir}, c_i^{CInd} and c_i^{Tot} will increase by factors of $EPSE_i^{Dir}\phi$, $EPSE_i^{CInd}\phi$ and $EPSE_i^{Tot}\phi$, respectively. Conversely, if each of the *ith* firm's K outputs decreases by a factor ϕ^{-1}, where the source of the decrease is the *ith* firm, composite spillovers to the firm and the combined sources of the *ith* firm and composite spillovers to the firm, c_i^{Dir}, c_i^{CInd} and c_i^{Tot} will decrease by factors of $\left(EPSE_i^{Dir}\phi\right)^{-1}$, $\left(EPSE_i^{CInd}\phi\right)^{-1}$ and $\left(EPSE_i^{Tot}\phi\right)^{-1}$, respectively. It therefore follows that $EPSE_i^{Dir}$, $EPSE_i^{CInd}$ and $EPSE_i^{Tot}$ $<$, $=$ or > 1 corresponds to increasing, constant or decreasing direct, composite indirect and total returns to scale along the specified portion of the relevant output expansion-path. Following the above approach for the spatial RSE measures, statistical inference for $EPSE_i^{Dir}$, $EPSE_i^{CInd}$ and $EPSE_i^{Tot}$ involves computing $1,000$ estimates.

Summarizing: if $EPSE_i^{Dir}$, $\left(EPSE_i^{CInd}\right)$, $\left[EPSE_i^{Tot}\right]$ is $<$, $=$ or > 1, as the *ith* firm moves along its direct (composite indirect) [total] output expansion-path from $\left((1-\lambda)\, y_i^{\mathbf{B}}, p_i^{\mathbf{B}}, t_i^{\mathbf{B}}\right)$ to $\left((1+\lambda)\, y_i^{\mathbf{B}}, p_i^{\mathbf{B}}, t_i^{\mathbf{B}}\right)$, c_i^{Dir}, $\left(c_i^{CInd}\right)$, $\left[c_i^{Tot}\right]$ will be rise by a smaller, the same or a larger proportion than the rise in y_i.

3. *Spatial Expansion-Path Subadditivity (EPSU)*

Berger *et al.* (1987) also present in a non-spatial setting $EPSU$, which is a combined measure of scale and product mix economies. To adapt their approach to the spatial case of $EPSU^{CInd}$ suppose the first points on two non-radial rays in the (y_i, p_i, t_i) space are $\left(y_i^{\mathbf{D}}, p_i^{\mathbf{D}}, t_i^{\mathbf{D}}\right)$ and $\left(y_i^{\mathbf{E}}, p_i^{\mathbf{E}}, t_i^{\mathbf{E}}\right)$. We then consider a further point $\left(y_i^{\mathbf{A}}, p_i^{\mathbf{A}}, t_i^{\mathbf{A}}\right)$ where $\left(y_i^{\mathbf{D}}, p_i^{\mathbf{D}}, t_i^{\mathbf{D}}\right) + \left(y_i^{\mathbf{E}}, p_i^{\mathbf{E}}, t_i^{\mathbf{E}}\right) = \left(y_i^{\mathbf{A}}, p_i^{\mathbf{A}}, t_i^{\mathbf{A}}\right)$. Here the three superscripts attached to the output vectors represent the corresponding points for three firms in the two-dimensional output space in figure 1, i.e., firm **A** produces the combined outputs of two smaller firms **D** and **E**, where all the outputs are the result of spillovers to the firms.

Using c_i^{CInd} from Eq. 6, $EPSU^{CInd}$ is calculated as follows.

$$EPSU_i^{CInd} = \frac{c_i^{CInd}\left(y_i^{\mathbf{D}}, p_i^{\mathbf{D}}, t_i^{\mathbf{D}}\right) + c_i^{CInd}\left(y_i^{\mathbf{E}}, p_i^{\mathbf{E}}, t_i^{\mathbf{E}}\right) - c_i^{CInd}\left(y_i^{\mathbf{A}}, p_i^{\mathbf{A}}, t_i^{\mathbf{A}}\right)}{c_i^{CInd}\left(y_i^{\mathbf{A}}, p_i^{\mathbf{A}}, t_i^{\mathbf{A}}\right)}. \tag{10}$$

The equations for $EPSU_i^{Dir}$, $EPSU_i^{CInd}$ and $EPSU_i^{Tot}$ have the same form so one can easily compute $EPSU_i^{Dir}$ and $EPSU_i^{Tot}$ by adapting Eq. 10. $EPSU_i^{Dir}$, $EPSU_i^{CInd}$ and $EPSU_i^{Tot}$ $<$, $=$ or > 0 corresponds to increasing, constant or decreasing direct, composite indirect and total returns to scale when a firm produces the sum of two smaller firms' output quantities, where the source of each output quantity is an individual firm, composite spillovers to the firm and the combined sources of an individual firm and composite spillovers to the firm, respectively. In the empirical application we conduct statistical inference for the estimates of $EPSU_i^{Dir}$, $EPSU_i^{CInd}$ and $EPSU_i^{Tot}$ by once again following the above approach for the spatial RSE measures .

As Wheelock and Wilson (2001) note in the non-spatial context, $EPSE$ and $EPSU$ are both combined measures of scale and product mix economies, but it is the latter that is more akin to a measure of returns to scope. This also applies to the $EPSU_i^{Dir}$, $EPSU_i^{CInd}$ and $EPSU_i^{Tot}$ measures we introduce. To illustrate, $EPSU_i^{CInd}$ is based on the change in the composite cost spillover to the *ith* firm, Δc_i^{CInd}, when the *ith* firm produces the sum of two smaller firms' output quantities, where the source of each output quantity is composite spillovers to the firm and the mix of the composite indirect spillover of outputs for the two smaller firms differs. $EPSE_i^{CInd}$, on the other hand, is based on Δc_i^{CInd} when there is only an *incremental* change in the *ith* firm's output levels that emanate from composite spillovers. Consequently, there is only an *incremental* change in the mix of the composite indirect spillover of outputs that gravitate to the *ith* firm. With $EPSU_i^{CInd}$ there are potentially large differences in the mix of the composite indirect spillover of outputs that gravitate to the larger firm and the two smaller firms.

Summarizing: if $EPSU_i^{Dir}$, $\left(EPSU_i^{CInd}\right)$, $\left[EPSU_i^{Tot}\right]$ is $<$, $=$ or $>$ 0 and the *ith* firm produces the combined output quantities of two smaller firms with different product mixes, where the source of each output quantity is the individual firm (composite spillovers to the firm) [the firm and composite spillovers to the firm combined], c_i^{Dir}, $\left(c_i^{CInd}\right)$, $\left[c_i^{Tot}\right]$ will rise by a smaller, the same or a larger proportion than the rise in y_i.

4 Application to U.S. Banks

4.1 Data and the Spatial Weights Matrices for the Different Regimes

Our data set is a rich balanced panel of annual observations for 387 large and medium-sized U.S. banks for the period 1998 – 2015. This is an interesting period as it includes the financial crisis as well as sufficiently long pre and post-crisis periods. Our sample is a balanced panel because we analyze continuously operating banks to focus on the spatial scale and product mix economies of the core group of surviving large and medium-sized institutions. Following Berger and Roman (2017) we classify a U.S. bank as large if its total assets in 2015 were greater than \$3 billion and medium-sized if in 2015 its total assets were between \$1 billion and \$3 billion. Based on this classification both bank size categories are well represented in our sample as there are 218 medium-sized banks and 169 large.[10]

The data for the variables is from the Call Reports sourced from the Federal Deposit Insurance Corporation (FDIC). Monetary values are deflated to 2005 prices using the consumer price index and the choice of output and input price variables is guided by the well-established intermediation approach to banking (Sealey and Lindley, 1977). See table 1 for descriptions of the outputs and input prices and summary statistics for the level variables. All the variables are logged and then mean adjusted so the first order direct, indirect and total output and normalized input price parameters can be interpreted as elasticities at the sample mean. The three outputs in our model specification, which reflect the lending and non-lending activities of banks, are total loans (y_1), total securities (y_2) and total non-interest income (y_3). There are three input prices that reflect the cost of fixed assets (p_1), labor (p_2) and deposits (p_3), where p_1 is the normalizing input price. Total operating cost (c) is the dependent variable and is the sum of the expenditures on the three inputs, where c is also normalized by p_1.

[10] Given we focus on medium-sized and large banks because their sufficiently large branch networks lead to a sufficient network overlap, and then distinguish between these bank size categories, size thresholds need to be used. There is not an industry accepted source for U.S. bank size thresholds so we use the thresholds from a leading recent paper in the banking literature (Berger and Roman, 2017). These thresholds are well-suited to our data set as they are for bank size in the final year of Berger and Roman's study period (2015), which is also the final year of our study period.

Table 1: Variable descriptions and summary statistics

Variable description	Model notation	Full Sample (387 banks) Mean	Std. Dev.
Total operating cost (in 000s of 2005 U.S. dollars): Sum of salaries, interest expenses on deposits and expenditure on fixed assets	c	394,215	2,333,162
Cost of fixed assets: Expenditure on fixed assets divided by the sum of the value of premises and fixed assets	p_1	0.320	0.596
Cost of labor (in 000s of 2005 U.S. dollars): Salaries divided by the total number of full-time equivalent employees	p_2	56.241	16.392
Cost of deposits: Interest expenses on deposits divided by total deposits	p_3	0.018	0.012
Loans: Net loans and leases (in 000s of 2005 U.S. dollars)	y_1	8,543,801	48,204,623
Total securities (in 000s of 2005 U.S. dollars)	y_2	2,781,599	17,894,416
Total non-interest income (in 000s of 2005 U.S. dollars)	y_3	304,672	1,979,639

We simultaneously use two specifications of $\mathbf{W}$ in our model. This involves splitting each bank's branch network into two types of branches- full service brick and mortar branches, and all other types of full and limited service branches. Each bank has both branch categories and so each bank features in both specifications of $\mathbf{W}$. We use $\mathbf{W_{BM}}$ to denote the full service brick and mortar based specification and $\mathbf{W_O}$ to denote the specification based on other branch types. Other types of branches can potentially cover up to twelve branch types, although seven is the most a bank has in our sample.[11]

The thinking behind our split of the banks' branch networks is that the split will capture different spatial linkages due to the different levels of centralization of activities across our two bank branch categories. There are a much large number of brick and mortar branches where similar general activities are highly decentralized across the branches. In contrast, there is a high degree of centralization of specialist activities in some of the other types of branches, which explains the relatively small number of full service cyber offices, limited service loan production offices and limited service consumer credit offices. With regard to the split of the bank branching networks, on average over our sample 88.4% are full service brick and mortar branches, which indicates that other branch types represent a relatively small share. An important related issue we examine is whether the degree of SAR dependence across overlapping brick and mortar networks is greater than across overlapping networks of other branches because of the dominance of brick and mortar branches in our sample. Alternatively, the SAR dependence could be greater across overlapping networks of other branches due to the higher degree of centralization of activities in these branches.

We construct $\mathbf{W_{BM}}$ and $\mathbf{W_O}$ in the same way using the following five steps.

(1) Begin with two matrices for each year in the sample and set all the cells on the main diagonal of a matrix to zero because a bank cannot be its own neighbor.

(2) For each state where the *ith* bank has the relevant type of branch (full service brick and mortar or other types of branches), we calculate the ratio of the number of *jth* bank branches to the

[11] For example, two of the twelve other types of branches are limited service loan production branches, which process loans and do not accept deposits, and limited service consumer credit branches, which only process consumer credit loans. For the full list of the twelve other branch types see https://www5.fdic.gov/sod/definitions.asp?systemform=soddnld3&helpitem=brsertyp&baritem=1.

number of *ith* bank branches. For the two matrices in each year we compute the non-zero off-diagonal elements by summing these ratios across the states where the *ith* bank has the relevant type of branch.[12]

(3) All other off-diagonal elements of the two matrices for each year are set to zero. These elements signify for the relevant type of branch that the *ith* and *jth* banks do not have overlapping branch networks.

(4) Average the annual matrices for each of the two branch categories from (3) to obtain two matrices for the sample.

(5) Normalize the elements of the average matrix for each of the two branch categories by dividing throughout by the largest element, which is referred to in the spatial literature as normalizing by the largest eigenvalue. The advantage of this normalization is that it preserves the information for the relevant type of branch on the absolute intensity of a bank's branch network vis-à-vis the overlapping branch network of another bank. This is because the normalization does not change the proportional relationship between the spatial weights. As a result, the spillovers between the banks are positively related to the absolute intensities.[13]

As $\mathbf{W_{BM}}$ and $\mathbf{W_O}$ are constructed using micro level geographical information on branch networks and the variables in the model are at the more aggregate level of the banking firm/bank holding company, based on parallels with firm level studies that include independent variables to reflect disaggregated plant level characteristics, it is reasonable to take $\mathbf{W_{BM}}$ and $\mathbf{W_O}$ to be exogenous.

As a final point on the data and spatial weights matrices, we note that there were a lot of mergers and acquisitions in the U.S. banking industry over our study period. Since our data set is a balanced panel and therefore contains only continuously operating medium-sized and large banks, this consolidation resulted in a non-negligible number of these banks increasing in size. Such size increases led to greater overlaps between banks' branch networks and, as a result, more interconnected banks. We in turn associate this with an increase in systemic risk, which was a key feature of the financial crisis. The greater overlap between banks' branch networks is inherent in the information about branch locations that we use to specify the spatial weights matrices. Moreover, the increase in interconnectedness and hence greater systemic risk are consistent with positive spatial correlation between banks' costs. This correlation will be reflected in the degree of spatial cost dependence across banks' overlapping brick and mortar branch networks and across their overlapping networks of other branch types (i.e., the estimates of δ_{BM} and δ_O).

4.2 Estimated Spatial Cost Model and the Elasticities

The fitted coefficients for our spatial Durbin cost function (SDCF) with two spatial regimes, $\mathbf{W_{BM}}$ and $\mathbf{W_O}$, are presented in table 2. Throughout our empirical analysis results that relate to $\mathbf{W_{BM}}$ and $\mathbf{W_O}$ (e.g., the parameters that are pre-multiplied by $\mathbf{W_{BM}}$ and $\mathbf{W_O}$ in table 2) apply to the

[12] As other types of branches represent a small share of bank branch networks, we calculate the above ratios at the rather aggregate state level to ensure there are overlapping networks of other branch types.

[13] Often the elements of a spatial weights matrix are normalized by the row sums. This is suitable when the elements are binary (e.g., elements that reflect contiguous geographical areas), which is very different from the case we consider. If we row-normalized our non-binary elements we would transform the information about the absolute intensities into relative intensities, which would make the spillovers difficult to interpret.

Table 2: Estimated spatial cost model with two simultaneous spatial regimes

	Model		Model		Model
	coeff		coeff		coeff
y_1	0.603^{***}	$\mathbf{W_{BM}}y_2$	0.171	$\mathbf{W_O}y_3$	-0.032
y_2	0.157^{***}	$\mathbf{W_{BM}}y_3$	-0.341^{*}	$\mathbf{W_O}p_2$	-0.470^{***}
y_3	0.177^{***}	$\mathbf{W_{BM}}p_2$	-0.108	$\mathbf{W_O}p_3$	0.012
p_2	0.535^{***}	$\mathbf{W_{BM}}p_3$	0.102	$\mathbf{W_O}y_1^2$	-0.013
p_3	0.368^{***}	$\mathbf{W_{BM}}y_1^2$	0.039	$\mathbf{W_O}y_2^2$	0.015
y_1^2	0.060^{***}	$\mathbf{W_{BM}}y_2^2$	-0.070	$\mathbf{W_O}y_3^2$	0.056
y_2^2	0.012^{***}	$\mathbf{W_{BM}}y_3^2$	-0.088^{*}	$\mathbf{W_O}y_1y_2$	0.056
y_3^2	0.046^{***}	$\mathbf{W_{BM}}y_1y_2$	-0.178	$\mathbf{W_O}y_1y_3$	-0.034
y_1y_2	-0.023^{***}	$\mathbf{W_{BM}}y_1y_3$	0.070	$\mathbf{W_O}y_2y_3$	-0.093^{*}
y_1y_3	-0.095^{***}	$\mathbf{W_{BM}}y_2y_3$	0.201^{*}	$\mathbf{W_O}p_2^2$	0.446^{***}
y_2y_3	0.006^{**}	$\mathbf{W_{BM}}p_2^2$	0.500^{**}	$\mathbf{W_O}p_3^2$	0.019^{*}
p_2^2	0.049^{***}	$\mathbf{W_{BM}}p_3^2$	-0.047^{***}	$\mathbf{W_O}p_2p_3$	-0.242^{***}
p_3^2	0.053^{***}	$\mathbf{W_{BM}}p_2p_3$	-0.161^{*}	$\mathbf{W_O}y_1p_2$	-0.424^{***}
p_2p_3	-0.091^{***}	$\mathbf{W_{BM}}y_1p_2$	-0.295	$\mathbf{W_O}y_1p_3$	0.083^{***}
y_1p_2	-0.051^{***}	$\mathbf{W_{BM}}y_1p_3$	-0.074	$\mathbf{W_O}y_2p_2$	-0.058
y_1p_3	0.070^{***}	$\mathbf{W_{BM}}y_2p_2$	0.242	$\mathbf{W_O}y_2p_3$	0.041^{*}
y_2p_2	0.012^{**}	$\mathbf{W_{BM}}y_2p_3$	-0.151^{***}	$\mathbf{W_O}y_3p_2$	0.486^{***}
y_2p_3	0.004^{*}	$\mathbf{W_{BM}}y_3p_2$	0.112	$\mathbf{W_O}y_3p_3$	-0.131^{***}
y_3p_2	0.029^{***}	$\mathbf{W_{BM}}y_3p_3$	0.137^{**}	$\mathbf{W_O}t$	0.006
y_3p_3	-0.062^{***}	$\mathbf{W_{BM}}t$	0.003	$\mathbf{W_O}t^2$	-0.002^{***}
t	0.001	$\mathbf{W_{BM}}t^2$	-0.002^{**}	δ_{BM}	0.214^{**}
t^2	0.002^{***}	$\mathbf{W_O}y_1$	-0.363^{***}	δ_O	0.268^{***}
$\mathbf{W_{BM}}y_1$	-0.017	$\mathbf{W_O}y_2$	0.135		

Notes: *, ** and *** denote statistical significance at the 5%, 1% and 0.1% levels, respectively.

same banks as each bank in our sample has both types of branches (i.e., full service brick and mortar branches and other types of full and limited service branches).[14]

It is now well-known that the spillover parameters from a model that contains one or more SAR variables are the indirect parameters. From Eqs. 4b and 4c we can see that the composite and decomposed indirect parameters depend on, among other things, the SAR parameters for the spatial regimes (i.e., the $\delta_1, ..., \delta_M$ parameters). Before we present and discuss the composite and decomposed indirect parameters, we note that even though a SAR coefficient is not an elasticity, it has an informative interpretation as it represents the degree of SAR dependence across the cross-sectional firms. From table 2 we can see that the estimates of δ_{BM} and δ_O are significant at the 1% level or less. In the context of the empirical spatial literature and in line with our expectations, the magnitudes of these estimates suggest non-negligible positive SAR cost dependence between banks with brick and mortar branches and other types of branches in the same state. As δ_{BM} is less than δ_O we conclude that the dominance of brick and mortar branches in our sample is a smaller source of SAR cost dependence than the high degree of centralization of activities in some of the other branch types. It is also evident from table 2 that a number of the coefficients on the local spatial variables are significant at the 5% level or less (e.g., the coefficients on $\mathbf{W_{BM}}y_3$, $\mathbf{W_O}y_1$ and $\mathbf{W_O}p_2$). These findings are supportive of our spatial Durbin model as opposed to a SAR model as the latter omits local spatial regressors.

In table 3 we present the direct, composite indirect and total parameters from our fitted SDCF, where for the moment we focus on the direct parameters. The first order direct output and input

[14] We thank an anonymous reviewer for suggesting that we point out that the results that relate to $\mathbf{W_{BM}}$ and $\mathbf{W_O}$ apply to the same banks.

Table 3: Estimated direct, composite indirect and total parameters

	Direct parameter	Composite indirect parameter	Total parameter
y_1	0.602^{***}	-0.085	0.517^{***}
y_2	0.157^{***}	0.172^{**}	0.329^{***}
y_3	0.177^{***}	-0.089	0.088
p_2	0.534^{***}	-0.192	0.342^{***}
p_3	0.368^{***}	0.123^{***}	0.491^{***}
y_1^2	0.060^{***}	0.017	0.076
y_2^2	0.012^{***}	-0.011	0.001
y_3^2	0.046^{***}	0.016	0.061^{**}
y_1y_2	-0.023^{***}	-0.027	-0.050
y_1y_3	-0.095^{***}	-0.018	-0.113^{*}
y_2y_3	0.006^{**}	0.010	0.016
p_2^2	0.050^{***}	0.431^{***}	0.481^{***}
p_3^2	0.053^{***}	0.008	0.060^{***}
p_2p_3	-0.092^{***}	-0.214^{***}	-0.305^{***}
y_1p_2	-0.052^{***}	-0.361^{***}	-0.413^{***}
y_1p_3	0.070^{***}	0.042	0.112^{***}
y_2p_2	0.012^{**}	0.047	0.059
y_2p_3	0.004^{*}	-0.024	-0.021
y_3p_2	0.030^{***}	0.333^{***}	0.363^{***}
y_3p_3	-0.062^{***}	-0.048^{*}	-0.110^{***}
t	0.001	0.005	0.006^{*}
t^2	0.002^{***}	-0.002^{***}	0.000

Notes: *, ** and *** denote statistical significance at the 5%, 1% and 0.1% levels, respectively.

price parameters indicate that the elasticities for these variables at the sample mean are positive and significant at the 0.1% level. Since the direct parameters can be interpreted in the same way as the parameters from the corresponding non-spatial cost function, the monotonicity and curvature properties of a non-spatial cost function also apply to the direct cost function (refer back to Eq. 5 for the general form of our direct translog cost function).

We conclude that our fitted model in table 2 is well-behaved as the resulting direct translog cost function in table 3 satisfies the monotonicity and concavity properties at virtually all of the data points.[15] We find that 98.1% of the sets of direct output and input price elasticities outside the sample mean satisfy the monotonicity property as they include only positive elasticities. In line with production theory, a cost function is concave in input prices (Kumbhakar and Lovell, 2000) if the input price Hessian is negative semi-definite (i.e., all the odd numbered principal minors are non-positive and all the even numbered ones are non-negative). Checking the sign pattern of the principal minors of a Hessian from a direct translog function involves following the approach in Glass *et al.* (2016), who apply the method for the non-spatial setting in Diewert and Wales (1987) to the spatial case. We observe that the reported direct translog cost function is almost exclusively concave in input prices over the sample, as 99.6% of the input price Hessians outside the sample mean are negative semi-definite and thus consistent with production theory.

We associate a negative and significant direct first order time parameter with annual Hicks neutral technical change for the sample average bank. The direct first order time parameter in table 3,

[15] We thank an anonymous reviewer for suggesting we investigate whether our reported direct translog cost function satisfies the monotonicity and concavity conditions outside the sample mean.

however, is positive, albeit only significant at the 10% level, so there are some signs that the model is picking up something other than Hicks neutral technical change. Based also on the positive direct second order time parameter, which is significant at the 0.1% level, there are signs that bank costs are characterized by a positive time trend that increases annually.

One can think of production theory as accounting for spillovers because the representative firm can be viewed as minimizing its costs that are under its control, having taken as given the cost spillovers. As the spillovers are taken as given, production theory does not posit the effect of a change in an input price (output) spillover to the representative firm on its cost. As a result, we do not consider further empirical monotonicity and curvature characteristics beyond the above analysis of the reported direct function that considers the portions of a bank's input prices, outputs and thus cost that are under the bank's control. We do not therefore explore the empirical monotonicity and curvature characteristics of the reported composite and decomposed indirect translog functions, which consider the portions of a bank's input prices, outputs and thus cost that are due to aggregate and disaggregated spillovers, respectively. Nor do we explore such characteristics of the reported total function as a part of the bank's input prices, outputs and thus cost that this function considers are due to aggregate spillovers.

Recall that a composite indirect parameter measures the cost spillover to the *ith* bank due to the spillover effect on one of the bank's own independent variables. We can see from table 3 that the composite indirect y_2 and p_3 parameters are significant and non-negligible, which further justifies adopting a spatial approach to cost modeling for U.S. banks. Both these composite parameters are positive because the correlations in the structural form of our model (Eq. 1) that are attributed to y_2 and p_3 in the reduced form of the model (Eq. 3) are positive. This is evident because the two SAR coefficients that capture the spatial correlation between banks' costs and the parameters that capture the spatial correlations between the effects of a bank's y_2 and p_3 variables on its cost (i.e., the $\mathbf{W_{BM}}y_2$, $\mathbf{W_O}y_2$, $\mathbf{W_{BM}}p_3$ and $\mathbf{W_O}p_3$ parameters) are all positive (see table 2). Both SAR parameters are significant, whereas the coefficients on the spatial lags of y_2 and p_3 are not significant. This indicates that the significant composite indirect y_2 and p_3 parameters are due to the SAR parameters dominating the coefficients on the spatial lags of y_2 and p_3. As a result of the direct and composite indirect y_2 and p_3 parameters, the total parameters for these variables are also positive and significant.

The composite indirect y_1, y_3 and p_2 parameters in table 3 are not significant. This is due to the negative spatial correlations between the effects of these variables on a bank's cost (as captured by the negative coefficients on the spatial lags of these variables in table 2) cancelling out the effect of the significant positive SAR parameters. For y_1 and p_2 the total parameters in table 3 are positive and significant because in both cases the positive and significant direct parameter more than offsets the relatively small, negative and insignificant composite indirect parameter. In contrast, the total y_3 parameter resembles different aspects of its direct and composite indirect components as it is positive like the former and not significant like the latter.

As ultimately it is the spatial scale and product mix economies we are interested in, we shed more light on the output spillover effects in table 4 by presenting the decomposed indirect parameters for the *ith* bank's first order outputs, squared outputs and interaction terms containing an output, where we use x to denote any of these variables. A composite indirect parameter does not indicate the different sources of the spillover effect on x, so in table 4 we provide the following three-way decomposition of this effect. (i) An indirect parameter when the spillover effect that is attributed to x is collectively from the two SAR variables, $\mathbf{W_{BM}}c$ and $\mathbf{W_O}c$ (column 2).[16] (ii) An indirect parameter

[16] It is not feasible to further decompose (i) into two indirect parameters relating to the separate spillover effects that

Table 4: The decomposition of the composite indirect parameters relating to the outputs

x	Decomposed indirect (i) parameter relating to the attribution to x from $\mathbf{W_{BM}}c$ and $\mathbf{W_O}c$ collectively	Decomposed indirect (ii) parameter relating to the attribution to x from $\mathbf{W_{BM}}x$	Decomposed indirect (iii) parameter relating to the attribution to x from $\mathbf{W_O}x$
y_1	0.138***	−0.006	−0.217***
y_2	0.036***	0.056	0.080
y_3	0.041***	−0.112*	−0.018
y_1^2	0.014***	0.011	−0.008
y_2^2	0.003***	−0.023	0.009
y_3^2	0.011***	−0.028	0.033
y_1y_2	−0.005***	−0.056	0.034
y_1y_3	−0.022***	0.023	−0.019
y_2y_3	0.001*	0.064*	−0.056*
y_1p_2	−0.012***	−0.099	−0.250***
y_1p_3	0.016***	−0.023	0.048**
y_2p_2	0.003*	0.079	−0.035
y_2p_3	0.001*	−0.049***	0.024*
y_3p_2	0.007***	0.038	0.288***
y_3p_3	−0.014***	0.044**	−0.077***

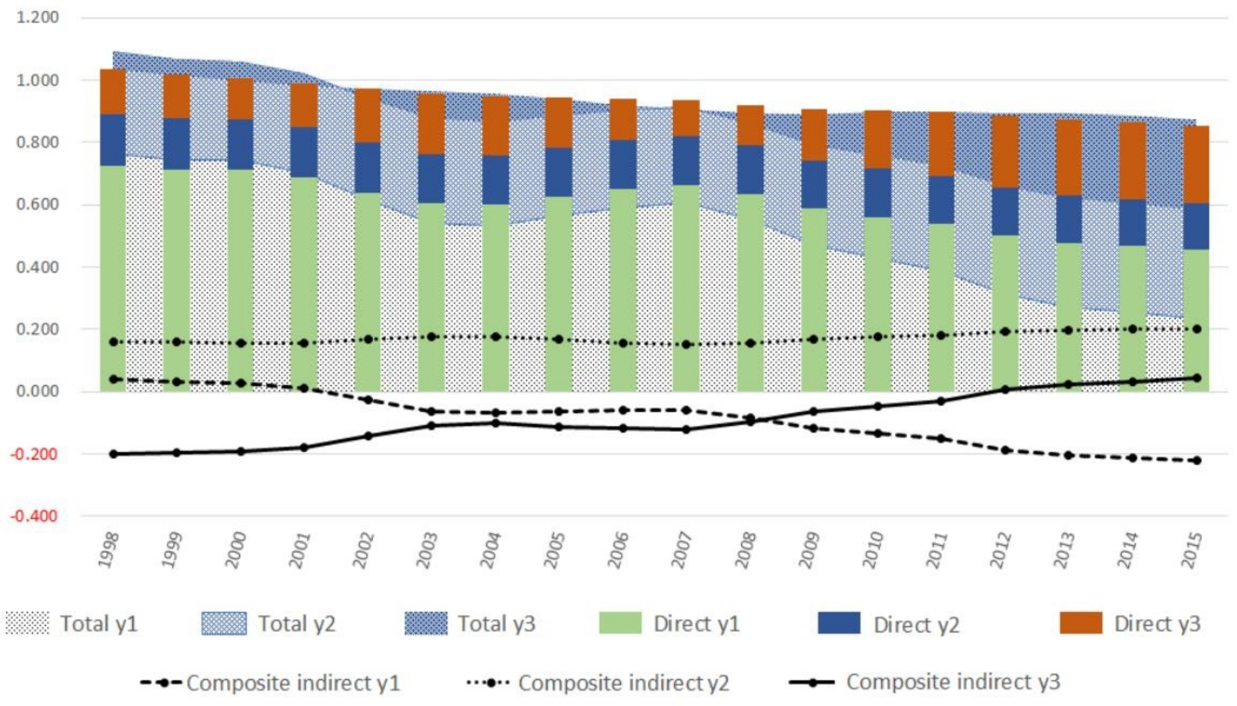
Notes: *, ** and *** denote statistical significance at the 5%, 1% and 0.1% levels, respectively.

when the spillover effect that is attributed to x is from $\mathbf{W_{BM}}x$ (column 3). (iii) An indirect parameter when the spillover effect that is attributed to x is from $\mathbf{W_O}x$ (column 4). The noteworthy features of table 4 are the positive and significant effects in column 2 for all the variables and the negative and significant effects in columns 3 and 4 for y_3 and y_1, respectively.

In figure 2 we present the average annual direct, composite indirect and total output elasticities to see how they evolve over the study period. We draw attention to the note directly below this figure which explains how we present the elasticities. From this figure we can see that the average direct elasticity for securities (y_2) is reasonably stable over the study period, whilst there are signs that the paths of the average direct elasticities for loans (y_1) and non-interest income (y_3) contrast. This is in line with the change in the nature of U.S. banking that is reported in the literature because, as Clark and Siems (2002) note, there was a fall in the commercial bank share of total U.S. financial intermediation, while there was an increase in commercial banks' non-interest income, which is heavily influenced by off-balance sheet activities. To illustrate, Clark and Siems report that the ratio of non-interest income to total income increased from 19% in the late 1970s to nearly 46% in 1999. Moreover, Lozano-Vivas and Pasiouras (2014) undertake a cross-country analysis of banks (1999 – 2006) and find that, on average, banks in major advanced countries (including the U.S.) have a higher ratio of non-interest income to total income.

We can also see from figure 2 that the average composite indirect elasticity for y_2 is positive, non-negligible and quite stable over the study period, whereas the average annual composite indirect elasticities for y_1 and y_3 are trending downwards and upwards. We can therefore conclude that, on average, the relationships between a bank's cost and the spillover effects on its y_1 and y_3 variables have changed considerably over the study period. To illustrate, due to the downward trend in the composite indirect elasticity for y_1, by the end of the study period we find that a marginal increase in the spillover effect on a bank's loans will lead to a marked fall in its cost. This is entirely plausible because based on the spatial literature (e.g., Kao and Bera, 2013, Boarnet and Glazer, 2002, and

are attributed to x from $\mathbf{W_{BM}}c$ and $\mathbf{W_O}c$ (Elhorst *et al.*, 2012).



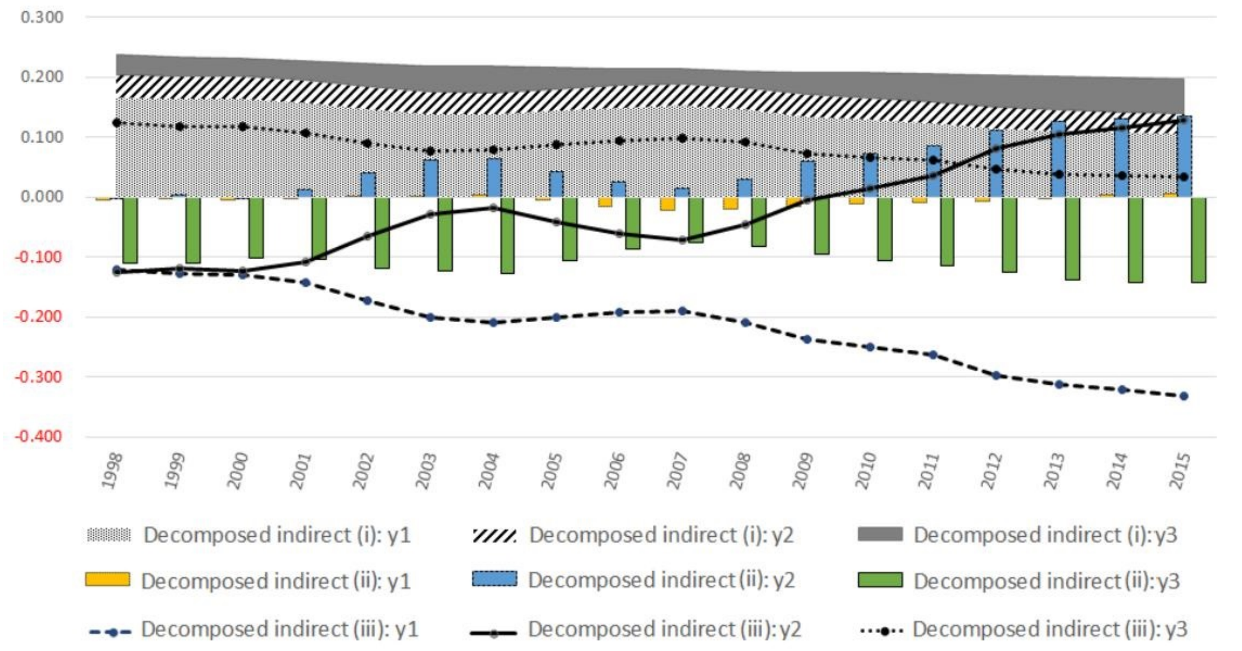
Note: The direct and total elasticities are presented in cumulative form from y1 through to y3, i.e., stacked bars and areas. To facilitate interpretation the composite indirect elasticities are not presented in cumulative form.

Figure 2: Average annual direct, composite indirect and total output elasticities

Garrett and Marsh, 2002) we would typically associate the increase in the magnitude of this negative relationship over a large part of the study period with more intense bank competition to make loans. This increase in competition fits with U.S. bank loan markets being fluid, which explains why studies have used data sets for different time periods to analyze if loan competition between U.S. banks changed over time (e.g., Berger *et al.*, 2004, 2007, and Bolt and Humphrey, 2015).

Following table 4, in figure 3 we present the three-way decomposition of the average annual composite indirect output elasticities (see (i)-(iii) above). Again we draw attention to the note below this figure which explains how we present these decomposed spillover elasticities. We can see from this figure that the average decomposed indirect (i) elasticity for y_2 exhibits a high degree of stability over the study period, whereas there are signs of a small trade-off between the same elasticities for y_1 and y_3. These elasticities for $y_1 - y_3$ are positive over the study period, which reflects the positive effects on cost from the spillovers that are attributed to the outputs from the two spatial lags of the dependent variable. Specifically, the decomposed indirect (i) elasticities for $y_1 - y_3$ are positive because the two positive SAR parameters are evidently capturing the positive spatial correlations between banks' costs. These spatial correlations are positive because banks that operate in the same markets and thus have overlapping branch networks are subject to common phenomena such as market growth and headline changes in local and regional economies.

Figure 3 also reveals that the average decomposed indirect (ii) elasticity for y_1 is very small over the study period. The same elasticity for y_3, despite a slight change around the 2008 financial crisis, is always non-negligible and negative. This reflects over our study period the negative effect on cost from the collective spillover effect that is attributed to y_3, y_3^2 and interaction terms that include y_3 from $\mathbf{W_{BM}}y_3$, $\mathbf{W_{BM}}y_3^2$ and the interaction terms that include y_3 and are pre-multiplied by $\mathbf{W_{BM}}$,



Note: The decomposed indirect (i) elasticities are presented in cumulative form from y1 through to y3, i.e., stacked areas. To facilitate interpretation the decomposed indirect (ii) and (iii) elasticities are not presented in cumulative form.

Figure 3: Average annual decomposed indirect output elasticities

respectively. See, for example, the large negative and significant coefficient on $\mathbf{W_{BM}}y_3$ in table 2. Specifically, the decomposed indirect (ii) y_3 elasticity over our study period is negative because collectively the $\mathbf{W_{BM}}y_3$, $\mathbf{W_{BM}}y_3^2$ and the interaction terms that include y_3 that are pre-multiplied by $\mathbf{W_{BM}}$ are capturing the negative spatial correlation between the collective effect of y_3, y_3^2 and the interaction terms that include y_3 on a bank's cost across the networks of brick and mortar branches. This spatial correlation is negative as banks that operate in the same markets via brick and mortar branches and thus have overlapping brick and mortar branch networks are evidently competing for non-interest income (y_3) (i.e., fees and charges).

The average decomposed (ii) indirect elasticity for y_2 goes from being very small in the early part of our study period to positive non-negligible levels in later years, which is again despite a dip around the crisis. In contrast to a decomposed (i) indirect output elasticity, which, as was noted above, relates to the spatial correlation between banks' costs across the two branch networks, decomposed (ii) and (iii) indirect output elasticities capture particular output phenomena (e.g., the aforementioned competition for non-interest income). Along the same lines, we suggest that the rise in the average decomposed (ii) indirect elasticity for y_2 (securities) in the second half of the study period is capturing the increase in the positive spillover effect on a bank's securities (and hence its cost) due to the increase in securities across the industry (He *et al.*, 2010; Langfield and Pagano, 2015).

$\mathbf{W_O}$ captures the overlap of banks' networks of other branch types. Recall that for some of these other types of branches we highlighted the high degree of centralization of particular activities in a branch vis-à-vis a brick and mortar branch. From figure 3 it is evident that the annual average decomposed indirect (iii) elasticity for y_1 (loans) is always negative and non-negligible, and tends to increase in magnitude over the study period. This is conceivably because other types of branches that specialize in loans (e.g., limited service loan production offices) are capturing the effect of the increases in bank competition to make loans that we asserted was driving the evolution of the composite indirect elasticity for loans in figure 2.

Table 5: Average spatial ray-scale economies for the full sample and large and medium-sized banks

Measure of RSE	Full sample	Large banks	Medium-sized banks
Average direct RSE, RSE^{Dir}	0.936^{*}	0.824^{*}	1.024^{*}
Average composite indirect RSE, RSE^{CInd}	-0.002_{a}	-0.045_{a}	0.032_{a}
Average total RSE, RSE^{Tot}	0.934^{*}	0.779^{*}	1.055^{*}
Average decomposed indirect (i) RSE, $RSE^{DInd}_{(i)}$	0.215^{*}_{a}	0.189^{*}_{a}	0.235^{*}_{a}
Average decomposed indirect (ii) RSE, $RSE^{DInd}_{(ii)}$	-0.061_{a}	-0.025_{a}	-0.089_{a}
Average decomposed indirect (iii) RSE, $RSE^{DInd}_{(iii)}$	-0.156^{*}_{a}	-0.209^{*}_{a}	-0.114^{*}_{a}

Notes: A bank is classified as large (medium-sized) if it has total assets greater than \$3 billion in 2015 (inbetween \$1 billion and \$3 billion in 2015) (Berger and Roman, 2017). At the 5% level * denotes significantly different from zero and a and b denote significantly less (greater) than or greater (less) than +1 (-1) for positive (negative) returns, respectively

4.3 Estimates of the Spatial Economies and the Policy Implications

For clarity we follow the non-spatial analyses by Wheelock and Wilson (2001; 2012) and discuss our estimates of the three different sets of spatial economies separately. We then pull together our key empirical findings from these sets to suggest the policy implications.

1. *Results and Analysis: Spatial Ray-Scale Economies (RSE)*

In table 5 we present for the full sample and two subsamples of large and medium-sized banks average estimates of RSE^{Dir}, RSE^{CInd}, RSE^{Tot} and the three measures of RSE^{DInd}, where $DInd$ denotes decomposed indirect. Each of the three reported estimates of RSE^{Dir} is not significantly less than or greater than 1, which points to constant RSE^{Dir}. For large banks the magnitude of the average estimate of RSE^{Dir} is some way below 1, which is consistent with the increase in the size of the largest banks since the financial crisis. 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 and two others (Bank of America and JPMorgan Chase) had more than \$1 trillion in assets. Compare this to the end of 2015 when the assets of the largest company (JPMorgan Chase) had increased to \$2.35 trillion with three others having more than \$1.7 trillion in assets.

The average estimate of RSE^{Dir} for large banks is not significantly different from constant returns due to the rather large standard error. This is what we would expect for large banks as there is a lot of variation in size in this sub-sample with some of the smaller large banks being much closer in size to medium-sized banks than very large banks, where for medium-sized banks the magnitude of the average estimate of RSE^{Dir} is only slightly above constant returns. As the direct parameters from a spatial model are akin to standard own parameters from a non-spatial model there is a close resemblance between direct and own returns to scale. It is not surprising therefore to find that our constant RSE^{Dir} results are in line with some of the results for own RSE from the non-spatial studies of U.S. banks by Wheelock and Wilson (2001; 2012).

We can see from table 5 that the three reported estimates of average composite spillover returns to scale to a bank (RSE^{CInd}) are not large and not significantly different from zero. This indicates that there are, on average, no cost implications for a bank from an equiproportional change in its output levels that are attributable to spillovers from other banks. Although we would have expected the three reported average estimates of RSE^{CInd} to suggest a non-negligible and significant change in the cost of a bank when there is an equiproportional change in its output levels due to spillovers, our subsequent findings, which we briefly summarize here and cover in more detail in due course, indicate that the relationship between a bank's cost and its output levels due to spillovers is more complex than this. This is for two reasons, first, although we report zero RSE^{CInd} estimates, closer analysis reveals that they are made up of significant non-negligible positive and negative decomposed spillover returns to scale (RSE^{DInd}) that cancel one another out. Second, we find that composite indirect scale economies are sensitive to the nature of the change in the output levels attributable to spillovers. To illustrate, whereas the zero RSE^{CInd} we report relate to a change in a bank's output levels due to spillovers along the radial ray, when we consider an incremental change in these output levels along the bank's non-radial expansion-path we observe $EPSE^{CInd}$ that are positive, non-negligible and significant, which is more in line with our expectations.

When the reported estimates of average RSE^{Dir} and RSE^{CInd} are summed to obtain RSE^{Tot} there is of course some similarity between the magnitudes of the corresponding average estimates of RSE^{Dir} and RSE^{Tot}. As is the case for the reported RSE^{Dir}, the average estimates of RSE^{Tot} for the full sample and large and medium-sized banks are not significantly less than or greater than 1, which suggests constant RSE^{Tot}. This indicates that, on average, if there is an equiproportional change in a bank's outputs levels that are under its control and its output levels due to spillovers, there will be the same proportionate change in the bank's cost.

In terms of the components of each of the insignificant RSE^{CInd}, we can see from table 5 that $RSE^{DInd}_{(i)}$ and $RSE^{DInd}_{(iii)}$ are significantly different from zero; $RSE^{DInd}_{(ii)}$ is insignificant; and $RSE^{DInd}_{(i)}$ and $RSE^{DInd}_{(iii)}$ are significantly less than or greater than $+1$ and -1, respectively. Since the decomposed indirect (i) elasticities for $y_1 - y_3$ are used to calculate $RSE^{DInd}_{(i)}$, the explanation of $RSE^{DInd}_{(i)}$ mirrors that for these elasticities. Accordingly, we conclude that the positive $RSE^{DInd}_{(i)}$ are capturing the positive spatial correlation between banks' scale economies because of the common phenomena that exist between banks that operate in the same markets.

Turning to the interpretation of $RSE^{DInd}_{(i)} - RSE^{DInd}_{(iii)}$. The estimates of $RSE^{DInd}_{(i)}$ in table 5 indicate that if there is an equiproportional increase in the spillover effect on a bank's outputs (i.e., on vector y) that emanates from a combined increase in $\mathbf{W_{BM}}c$ and $\mathbf{W_O}c$, $c^{DInd}_{(i)}$ will rise by smaller proportion than y. Conversely, from the negative estimates of $RSE^{DInd}_{(iii)}$ we can conclude that if due to an increase in the spillover effect on y that emanates from an increase in $\mathbf{W_O}y$ there is an equiproportional decrease in a bank's outputs, which we attribute to the effects of competition, $c^{DInd}_{(iii)}$ will decline by a smaller proportion than y. Interestingly, it is evident that we obtain insignificant average RSE^{CInd} for the full sample and large and medium-sized banks because in all three cases the positive and significant $RSE^{DInd}_{(i)}$ is cancelled out by the negative and insignificant $RSE^{DInd}_{(ii)}$ and the negative and significant $RSE^{DInd}_{(iii)}$.

In figure 4 we present the sample average annual RSE^{Dir}, RSE^{CInd} and RSE^{Tot} together with the 95% confidence intervals. The general conclusion from this figure is that the evolution of average

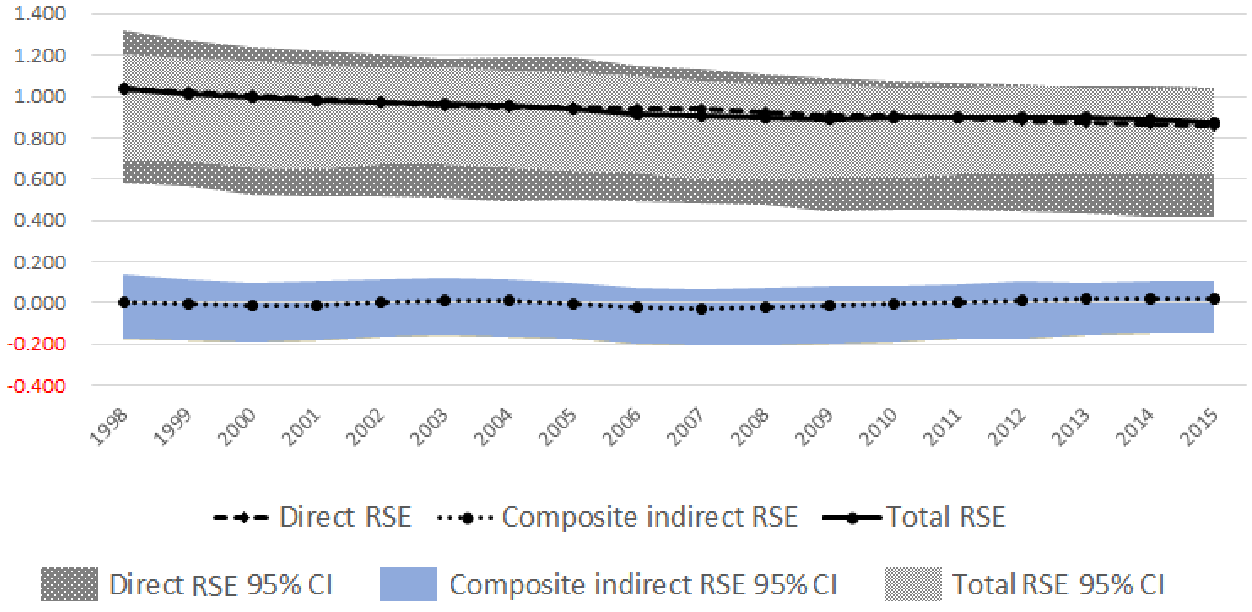


Figure 4: Average annual direct, composite indirect and total ray-scale economies

annual RSE^{Dir}, RSE^{CInd} and RSE^{Tot} reflect what we observed in table 5. This is because average annual RSE^{CInd} is always not significantly different from zero and, as a result, average annual RSE^{Tot} mirrors its RSE^{Dir} component. We can also see from figure 4 that average annual RSE^{Dir} and RSE^{Tot} are on downward trends and by the end of the study period both are close to being significantly less than 1. This suggests that, on average, banks are moving towards becoming smaller than their minimum efficient size.

An interesting issue that emerges from figure 4 is the sources of the insignificant annual RSE^{CInd} over the study period. To examine these sources in figure 5 we present the decomposition of the annual RSE^{CInd} into its three constituent parts ($RSE^{DInd}_{(i)} - RSE^{DInd}_{(iii)}$). From this figure we can see that annual $RSE^{DInd}_{(i)} - RSE^{DInd}_{(iii)}$ closely resemble the results for the averages of these measures over the study period (see table 5). This is evident as we always observe annual $RSE^{DInd}_{(i)}$ and $RSE^{DInd}_{(iii)}$ that are significantly different from zero; an insignificant annual $RSE^{DInd}_{(ii)}$; and annual $RSE^{DInd}_{(i)}$ and $RSE^{DInd}_{(iii)}$ that are significantly less than or greater than +1 and -1, respectively. Given the similarity between an annual RSE^{DInd} measure and its average over the study period the conclusions we reach are similar. We conclude that the steadily declining positive annual $RSE^{DInd}_{(i)}$ in figure 5 is due to a declining positive spatial correlation between the scale economies of banks that operate in the same markets because the common phenomena between these banks is having a smaller effect. We also attribute the increasingly negative annual $RSE^{DInd}_{(iii)}$ to progressively more intense competition over the study period.

Looking ahead, from figure 4 insignificant annual RSE^{CInd} seems to be an interesting persistent feature of our sample of U.S. banks, which is in no way an artifact of our model specification or estimator. Given the interesting trends of the three components of annual RSE^{CInd} over the study period, researchers need to continue checking whether annual RSE^{CInd} becomes significantly different from zero and, if so, check which components are driving the change.

2. *Results and Analysis: Spatial Expansion-Path Scale Economies (EPSE)*

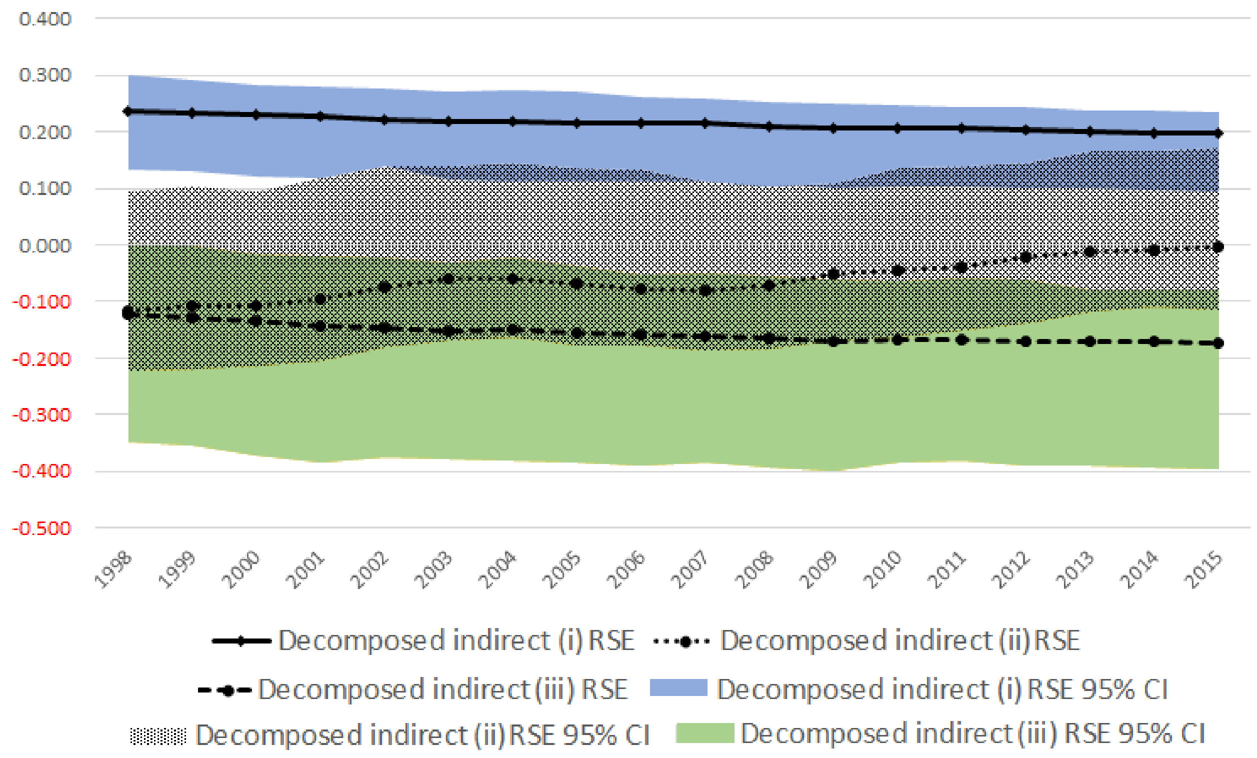


Figure 5: Average annual decomposed indirect ray-scale economies

In table 6 we present for the full sample and two subsamples of large and medium-sized banks average estimates of $EPSE^{Dir}$, $EPSE^{CInd}$, $EPSE^{Tot}$ and the three measures of $EPSE^{DInd}$. The average estimates of $EPSE^{Dir}$ for the full sample and large banks are not significantly different from 1, which points to constant returns and is in line with our RSE^{Dir} estimates. In the same way as RSE^{Dir} and own RSE from a non-spatial study have a close resemblance, finding evidence of constant $EPSE^{Dir}$ is in line with some of the own $EPSE$ results for U.S. banks that Wheelock and Wilson (2001; 2012) report. There have though been claims that large U.S. banks are "too big", which although may be valid for an individual large bank on a case-by-case basis, our average $EPSE^{Dir}$ estimate for large banks suggests that this is not a typical feature of this bank size category as we find that the size of the average large bank (and also the average bank in our full sample) is at its minimum efficient level. For medium-sized banks, in contrast to the constant average RSE^{Dir}, the average $EPSE^{Dir}$ suggests significant decreasing returns, although the magnitude of these returns is only marginally above 1. If anything, this suggests that it is medium-sized banks where the average bank size is above its minimum efficient scale.

Recall that summing the direct output elasticities gives the constant RSE^{Dir} estimates in table 5. We have seen that these RSE^{Dir} results are in line with the constant $EPSE^{Dir}$ estimates in table 6 as the numerator and denominator in the calculation of these $EPSE^{Dir}$ estimates are evidently not too different. By summing the composite indirect output elasticities we also obtained the RSE^{CInd} estimates in table 5, which are not significantly different from zero. In table 6 we provide a different insight into returns to scale spillovers as we report constant $EPSE^{CInd}$. On average, this suggests that an increase along a bank's composite indirect output expansion-path from 95% of its mean vector of outputs attributable to composite spillovers to 105% will lead to the bank's cost also rising by 10%. Our findings therefore suggest along the relevant radial ray and non-radial expansion-path that there is a much bigger difference between expansion of a bank's outputs that are attributable to composite

Table 6: Average spatial expansion-path scale economies for the full sample and large and medium-sized banks

Measure of EPSE	Full sample	Large banks	Medium-sized banks
Average direct EPSE, $EPSE^{Dir}$	1.020^*_{b}	1.000^*_{b}	1.040^*_{c}
Average composite indirect EPSE, $EPSE^{CInd}$	1.021^*_{b}	1.040^*_{b}	0.976^*_{b}
Average total EPSE, $EPSE^{Tot}$	1.026^*_{b}	0.999^*_{b}	1.056^*_{c}
Average decomposed indirect (i) EPSE, $EPSE^{DInd}_{(i)}$	1.020^*_{b}	1.000^*_{b}	1.039^*_{c}
Average decomposed indirect (ii) EPSE, $EPSE^{DInd}_{(ii)}$	0.972^*_{b}	1.018^*_{b}	0.926^*_{b}
Average decomposed indirect (iii) EPSE, $EPSE^{DInd}_{(iii)}$	1.022^*_{b}	1.041^*_{c}	1.002^*_{b}

Notes: A bank is classified as large (medium-sized) if it has total assets greater than \$3 billion in 2015 (inbetween \$1 billion and \$3 billion in 2015) (Berger and Roman, 2017). Based on 95% confidence intervals * denotes significantly different from zero and a, b and c denote significantly less than, equal to or greater than 1, respectively.

spillovers than there is between the expansion of the bank's standard type of outputs that are under its control. It is not entirely surprising though that there is greater similarity between RSE^{Dir} and $EPSE^{Dir}$ vis-à-vis the difference between RSE^{CInd} and $EPSE^{CInd}$. This is because underlying direct scale economies is the standard theoretical cost function that is monotonically increasing in a firm's output levels that are under its control, whereas there is no such theoretical relationship between a firm's cost and its output levels that are attributable to composite/decomposed spillovers, which are primarily, but not entirely, outside its control.

In contrast to RSE^{Tot} being the sum of RSE^{Dir} and RSE^{CInd}, $EPSE^{Dir}$ and $EPSE^{CInd}$ are ratios with different denominators so summing them does not give $EPSE^{Tot}$. From table 6 we can see that the estimates of $EPSE^{Tot}$ for the full sample and large banks are not significantly different from 1, which suggests constant returns along the two associated total output expansion-paths. For medium-sized banks the estimate of $EPSE^{Tot}$ is significantly greater than 1. The reported estimates of $EPSE^{Tot}$ are therefore in line with the corresponding $EPSE^{Dir}$. To elaborate on this recall that total output is output which is under a bank's control plus the composite spillovers from other banks to the bank's output. The $EPSE^{Tot}$ estimates therefore suggest that the average bank in the full sample and the average large bank are operating at their minimum efficient total output levels, and that the total output of the average medium-sized bank is greater than its minimum efficient level. We suggested that the magnitude of the $EPSE^{Dir}$ for the average medium-sized bank is marginally greater than 1, whereas the magnitude of the $EPSE^{Tot}$ for this bank can be viewed as being clearly above 1.

In line with what we have discussed about RSE^{Tot} and $EPSE^{Tot}$, the sum of $RSE^{DInd}_{(i)} - RSE^{DInd}_{(iii)}$ is RSE^{CInd} but $EPSE^{DInd}_{(i)} - EPSE^{DInd}_{(iii)}$ have different denominators so summing them does not yield $EPSE^{CInd}$. Each of the average $EPSE_{(i)}^{DInd} - EPSE_{(iii)}^{DInd}$ estimates for the full sample is not significantly different from 1, which is consistent with the corresponding $EPSE^{CInd}$. For large banks its the average $EPSE_{(i)}^{DInd}$ and $EPSE_{(ii)}^{DInd}$ that are not significantly different from 1 and in line with the corresponding $EPSE^{CInd}$, whilst the average $EPSE_{(iii)}^{DInd}$ is significantly greater than 1. For medium-sized banks it is the average $EPSE_{(i)}^{DInd}$ that is significantly greater than 1, whereas the average $EPSE_{(ii)}^{DInd}$ and $EPSE_{(iii)}^{DInd}$ are not significantly different from 1 and thus consistent with the corresponding $EPSE^{CInd}$.

3. *Results and Analysis: Spatial Expansion-Path Subadditivity (EPSU)*

For banks across consecutive deciles of our bank size distribution, where bank size is measured using total assets, in table 7 we present estimates of $EPSU^{Dir}$, $EPSU^{CInd}$, $EPSU^{Tot}$ and the three measures of $EPSU^{DInd}$. To illustrate, 1st decile/2nd decile $EPSU^{Dir}$, $EPSU^{CInd}$, $EPSU^{Tot}$ and $EPSU_{(i)}^{DInd} - EPSU_{(iii)}^{DInd}$ (see (a)-(d) below, respectively) measure the potential proportional difference in the direct, composite indirect, total and decomposed indirect (i)/(ii)/(iii) cost of an average 2nd decile bank with its output mix compared to the following.

(a) The combined direct costs (i.e., cost that is under the control of a bank and is net of the composite indirect cost spillover to the bank from the other banks in the sample) of two smaller banks with different output mixes. These two banks are the bank producing the average output vector across the 1st decile and the bank producing the difference between the average output vectors in the 2nd and 1st deciles.

(b) The combined composite indirect cost spillovers to the two smaller banks in (a).

(c) The combined total costs (i.e., direct cost of a bank plus the composite indirect cost spillover to the bank) of the two smaller banks in (a).

(d) The combined decomposed indirect (i)/(ii)/(iii) cost spillovers to the two smaller banks in (a).

As was also the case for the above spatial $EPSE$ results, $EPSU^{Dir}$, $EPSU^{CInd}$ and $EPSU_{(i)}^{DInd} - EPSU_{(iii)}^{DInd}$ are ratios with different denominators, so summing $EPSU^{Dir}$ and $EPSU^{CInd}$ does not yield $EPSU^{Tot}$, and $EPSU^{CInd}$ is not the sum of $EPSU_{(i)}^{DInd} - EPSU_{(iii)}^{DInd}$. To explain the interpretation of the estimates in table 7 consider for the medium-sized distribution: the significant negative $EPSU^{Dir}$ for the 1st decile/2nd decile; the significant positive $EPSU_{(i)}^{DInd}$ for the 6th decile/7th decile; and the $EPSU^{Tot}$ for the 2nd decile/3rd decile, which is not significantly different from zero. These estimates suggest that the direct cost of the average 2nd decile bank, the decomposed indirect (i) cost of the average 7th decile bank and the total cost of the average 3rd decile bank are higher than, lower than and not significantly different from the sum of the relevant costs of the two smaller banks.

There are two prominent features of table 7. First, from the $EPSU^{Dir}$, $EPSU^{CInd}$ and $EPSU^{Tot}$ results we can see that it is the $EPSU^{Dir}$ estimates which are key as all but one is significant, whilst all but one of the $EPSU^{CInd}$ estimates and all of the $EPSU^{Tot}$ results are not significant. Second, there are sequences across consecutive deciles of significant positive/negative estimates of $EPSU^{Dir}$, $EPSU_{(i)}^{DInd}$ and $EPSU_{(iii)}^{DInd}$. Over our entire bank size distribution there are two $EPSU^{Dir}$, $EPSU_{(i)}^{DInd}$ and $EPSU_{(iii)}^{DInd}$ cycles, where a cycle comprises sequences of negative and then positive estimates.

Table 7: Estimates of spatial expansion-path subadditivity for different bank sizes

Bank size		Direct EPSU, $EPSU^{Dir}$	Composite indirect EPSU, $EPSU^{CInd}$	Total EPSU, $EPSU^{Tot}$	Decomposed indirect (i) EPSU, $EPSU^{DInd}_{(i)}$	Decomposed indirect (ii) EPSU, $EPSU^{DInd}_{(ii)}$	Decomposed indirect (iii) EPSU, $EPSU^{DInd}_{(iii)}$
Medium	1st decile/2nd decile	−0.030*	−1.000	0.012	−0.030*	−0.127	−0.167*
	2nd decile/3rd decile	−0.043*	−0.978	0.018	−0.043*	−0.145	−0.246*
	3rd decile/4th decile	−0.068*	−0.742	0.030	−0.068*	−0.150	−0.387*
	4th decile/5th decile	−0.128*	−0.577	0.071	−0.128*	−0.152	−0.688*
	5th decile/6th decile	−0.795*	−0.451*	−0.213	−0.793*	−0.151	−2.113*
	6th decile/7th decile	0.180*	−0.394	−0.043	0.181*	−0.152	1.371
	7th decile/8th decile	0.072*	−0.304	−0.021	0.073*	−0.141	0.511
	8th decile/9th decile	0.041*	−0.242	−0.013	0.041*	−0.128	0.273
	9th decile/10th decile	0.008*	−0.031	−0.002	0.008*	0.083	−0.078*
Medium/Large	10th decile/1st decile	−0.029*	−5.566	0.011	−0.029*	−0.277	−0.144*
Large	1st decile/2nd decile	−0.061*	1.873	0.023	−0.062*	2.391	−0.303*
	2nd decile/3rd decile	−0.251*	1.557	0.079	−0.252*	0.352	−0.843
	3rd decile/4th decile	0.172*	1.614	−0.078	0.173*	0.280	1.267
	4th decile/5th decile	0.065*	1.568	−0.027	0.065*	0.200*	0.399*
	5th decile/6th decile	0.039*	1.647	−0.015	0.039*	0.143*	0.238*
	6th decile/7th decile	0.026*	1.268	−0.010	0.026*	0.108*	0.150*
	7th decile/8th decile	0.018*	0.595	−0.007	0.018*	0.076*	0.098*
	8th decile/9th decile	0.011*	0.385	−0.005	0.011*	0.048*	0.062*
	9th decile/10th decile	0.004	−0.019	−0.001	0.004	0.025*	0.077

Notes: A bank is classified as large (medium-sized) if it has total assets greater than $3 billion in 2015 (inbetween $1 billion and $3 billion in 2015) (Berger and Roman, 2017). * denotes significantly different from zero based on the 95% confidence interval.

We use our RSE and $EPSE$ results to inform the policy debate on whether there should be size caps on very large U.S. banks. We use these results as the basis for the policy implications rather than the $EPSU$ results because RSE and $EPSE$ relate to actual banks in the data space, whereas $EPSU$ considers amalgamations of hypothetical non-existent banks that may be far removed from actual banks (Wheelock and Wilson, 2001).[17] That said, it is still appropriate to include in this paper the method to calculate the spatial $EPSU$ measures together with our application of these measures to U.S. banks. This is because our $EPSU^{Dir}$ results have an interesting cyclical pattern over the bank size distribution, which can provide an indication to policy makers about particular bank sizes that may look to become larger to exploit scale economies. In particular, the positive and significant $EPSU^{Dir}$ results for some of the largest hypothetical banks suggest that there is an incentive for these hypothetical banks to become even larger to realize scale economies.

The policy implications from our results relate to the absence of any counteraction for large banks between: (i) the average RSE^{CInd} and RSE^{Dir}; and (ii) the average $EPSE^{CInd}$ and $EPSE^{Dir}$. The absence of counteraction in (i) and (ii), however, is for different reasons. In the case of (i), since the average RSE^{CInd} for large banks is zero the average ray-scale returns for large banks depends only on RSE^{Dir} and not also on the spillover of scale economies to a large bank from other banks. The average RSE^{Dir} for large banks points to constant returns to scale for the standard type of output levels that are under a bank's control, which suggests that large banks are using society's resources efficiently to provide their services. Size caps on very large banks would place downward pressure on the average RSE^{Dir} across the large bank size category, and if this leads to significant scale economies then large banks would be using society's resources inefficiently. Although in terms of the efficient use of society's resources our average RSE^{Dir} for large banks is not supportive of size caps on very large banks, there may be a case for size caps on these banks if they are TBTF and taking excessive risks as a result. This issue is very different to one we consider and is outside the scope of this paper.

Turning to discuss (ii) above, since the average $EPSE^{CInd}$ for large banks is not significantly different from 1 this suggests that, on average, there are constant returns for a large bank from the size spillovers to the bank from other banks. As the average $EPSE^{Dir}$ for large banks is also not significantly different from 1, size caps on very large banks would place downward pressure on both of the average $EPSE^{Dir}$ and $EPSE^{CInd}$ measures for large banks, where these measures relate to the output levels that are under a bank's control and the spillover effect on these output levels which is primarily, but not entirely, outside the control of the bank. If this downward pressure leads to significant scale economies for both measures, in contrast to (i) where the downward pressure from the size caps on the average RSE^{Dir} for large banks would represent a single source of inefficiency in the use of society's resources, the downward pressure on the average $EPSE^{Dir}$ and $EPSE^{CInd}$ for large banks would represent two such sources. Despite this difference between the number of potential sources, in terms of only the efficiency of the use of society's resources and in line with our policy implications from our average RSE^{Dir} for large banks, our average $EPSE^{Dir}$ and $EPSE^{CInd}$ for large banks are also not supportive of size caps on very large banks.

To illustrate for large banks the impact of the absence of counteraction between the average RSE^{CInd} and RSE^{Dir} and between the average $EPSE^{CInd}$ and $EPSE^{Dir}$, consider the following situation. Suppose for large banks the average $EPSE^{CInd}$ was 0.05 higher (i.e., increases to 1.09),

[17] Although we cannot of course be certain, this may be why in their two most recent papers on scale economies in U.S. banking Wheelock and Wilson (2012; 2018) do not calculate $EPSU$, but in an earlier paper (Wheelock and Wilson, 2001) they do.

where this increase is sufficient for the average $EPSE^{CInd}$ to go from constant returns to statistically significant decreasing returns. At the same time suppose the average $EPSE^{Dir}$ for large banks was 0.05 lower (i.e., declines to 0.95), where this decrease is sufficient for the average $EPSE^{Dir}$ to go from constant returns to statistically significant increasing returns. Note that to make out point we choose to adjust $EPSE^{CInd}$ and $EPSE^{Dir}$ because the adjustments do not represent very big departures from our $EPSE$ results for large banks. The point we make also applies to RSE^{CInd} and RSE^{Dir} although the changes we would need to make to these measures for large banks to make the same point would have to be much bigger and would therefore be less plausible, but this is not say that this would be the case for all applications. Moreover, as we observe zero average RSE^{CInd} for the full sample and large and medium-sized banks, the classification of the returns as increasing/constant/decreasing is redundant, although again this is not say that these returns would be zero in other applications. It is plausible in another application that there are increasing average RSE^{Dir} that are not too far below 1 and increasing average RSE^{CInd} that are well below 1 but significantly greater than 0. Even though the classification of the average RSE^{CInd} is clear the magnitude of these returns is important as they may be large enough when the average RSE^{Dir} and RSE^{CInd} are added together to give constant or increasing average RSE^{Tot}.

There would be counteraction between the policy implications from the adjusted values of the average $EPSE^{CInd}$ and $EPSE^{Dir}$ for large banks. This is because in terms of the average output vector across the large banks, on one hand, the average $EPSE^{CInd}$ points to this output vector being too large due to composite indirect diseconomies of scale, whilst on the other, the average $EPSE^{Dir}$ suggests that the vector is too small as there are direct economies of scale. As the composite indirect diseconomies of scale are greater than the direct economies, from only a cost perspective this would indicate that, on average, large banks are too big and should be sub-optimally smaller. The upshot is that due to the possibility of counteraction it is important to calculate composite indirect scale economies and to classify these returns as increasing/constant/decreasing.[18]

5 Concluding Remarks and Further Work

This paper sets out the methodology to extend well-established non-spatial measures of scale and product mix economies that are internal to a bank to the spatial case. We consider the interesting situation where banks simultaneously belong to multiple spillover regimes, which more generally is typically what we observe for firms. Our approach introduces sets of spatial measures of RSE, $EPSE$ and $EPSU$. Using some terminology from the spatial literature each of these sets comprises direct, composite and decomposed indirect and total measures.

The key findings from our empirical application to a sample of large and medium-sized U.S. banks ($1998-2015$) relates to the large difference between the composite indirect RSE and $EPSE$ vis-à-vis the similarity between the corresponding direct estimates. These direct measures are akin to standard internal scale economies from a non-spatial model and are therefore interpreted in the same way. The composite indirect measures relate to the cost spillover to a firm when there is a change in its output levels that are attributable to the composite spillovers from other firms in the sample.

For our full sample and subsamples of large and medium-sized banks we often observe constant direct RSE and $EPSE$, while we also report zero composite indirect RSE and constant composite indirect $EPSE$. There are not big implications therefore from only focusing on direct RSE or direct

[18] We thank an anonymous reviewer for suggesting that we emphasize why this classification of composite indirect scale economies is informative.

$EPSE$, but this is certainly not the case for the composite indirect RSE and $EPSE$, which is why it is important to calculate both these measures. This is because along the relevant radial ray and non-radial expansion-path there is a much bigger difference between a change in a bank's output levels that are attributable to spillovers and are thus primarily, but not entirely, outside its control, than there is between a change in its standard type of output levels that are under its control. In other words, there is a much bigger difference between the composite indirect radial ray and the composite indirect expansion-path than there is between the direct radial ray and direct expansion-path. This should not come as a great surprise because underlying direct/internal scale economies is the textbook theoretical cost function that is monotonically increasing in a firm's output levels that are under its control, whereas there is no such theoretical relationship between a firm's cost and its output levels that are attributable to spillovers, as these outputs are primarily, but not entirely, outside its control.

As a result of our paper there is a lot of scope for further banking applications of spatial economies because the underlying idea of cost spillovers between banks is quite intuitive. As our paper is the first of its type our approach focuses on providing the necessary comprehensive coverage of the methodology. Further applications can be simpler and more applied and policy focused than the application we provide because as a result of our paper further applications need not revisit the methodology in such detail. Simpler further applications could consider just one spatial network rather than multiple ones; focus on a subset of the spatial economies we introduce (e.g., the non-spatial analyses by Wheelock and Wilson (2012; 2018) focus on RSE and $EPSE$); and/or overlook the decomposed indirect spatial economies by considering only the direct, composite indirect and total returns.

References

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

Anselin, L. (2003): 'Spatial externalities, spatial multipliers and spatial econometrics'. *International Regional Science Review*, vol. 26, pp. 153-166.

Asmild, M. and K. Matthews (2012): 'Multi-directional efficiency analysis of efficiency patterns in Chinese banks 1997-2008'. *European Journal of Operational Research*, vol. 219, pp. 434-444.

Asmild, M. and M. Zhu (2016): 'Controlling for the use of extreme weights in bank efficiency assessments during the financial crisis'. *European Journal of Operational Research*, vol. 251, pp. 999-1015.

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

Bera, A. K., O. Dǒgan and S. Taşpinar (2018): 'Simple tests for endogeneity of spatial weights matrices'. *Regional Science and Urban Economics*, vol. 69, pp. 130-142.

Berger, A. N., A. Demirgűç-Kunt, R. Levine and J. G. Haubrich (2004): 'Bank concentration and competition: An evolution in the making'. *Journal of Money, Credit and Banking*, vol. 36, pp. 433-451.

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

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

Berger, A. N., R. J. Rosen and G. F. Udell (2007): 'Does market size structure affect competition? The case of small business lending'. *Journal of Banking and Finance*, vol. 31, pp. 11-33.

Bibi, U., H. O. Balli, C. D. Matthews and D. W. L. Tripe (2018): 'Impact of gender and governance on microfinance efficiency'. *Journal of International Financial Markets, Institutions and Money*, vol. 53, pp. 307-319.

Boarnet, M. G. and A. Glazer (2002): 'Federal grants and yardstick competition'. *Journal of Urban Economics*, vol. 52, pp. 53-64.

Bolt, W. and D. Humphrey (2015): 'Assessing bank competition for consumer loans'. *Journal of Banking and Finance*, vol. 61, pp. 127-141.

Christensen, L. R., D. W. Jorgenson and L. J. Lau (1973): 'Transcendental logarithmic production frontiers'. *Review of Economics and Statistics*, vol. 55, pp. 28-45.

Clark, J. A. and T. F. Siems (2002): 'X-efficiency in banking: Looking beyond the balance sheet'. *Journal of Money, Credit and Banking*, vol. 34, pp. 987-1013.

Conley, T. G. (1999): 'GMM estimation with cross sectional dependence'. *Journal of Econometrics*, vol. 92, pp. 1-45.

Conley, T. G. and G. Topa (2002): ‘Socio-economic distance and spatial patterns in unemployment’. *Journal of Applied Econometrics*, vol. 17, pp. 303-327.

Diewert, W. E. and T. J. Wales (1987): ‘Flexible functional forms and global curvature conditions’. *Econometrica*, vol. 55, pp. 43-68.

Economist (2008): *Idea- Economies of scale and scope.* Oct. 20th, 2008.

Elhorst, J. P. (2009): *Spatial panel data models.* In the *Handbook of Applied Spatial Analysis,* Fischer, M. M., and A. Getis (Eds). New York: Springer.

Elhorst, J. P. and S. Fréret (2009): ‘Evidence of political yardstick competition in France using a two-regime spatial Durbin model with fixed effects’. *Journal of Regional Science*, vol. 49, pp. 931-951.

Elhorst, J. P., D., J. Lacombe and G. Piras (2012): ‘On model specification and parameter space definitions in higher order spatial econometric models’. *Regional Science and Urban Economics*, vol. 42, pp. 211-220.

Fisher, R. W. and H. Rosenblum (2012): ‘Vanquishing too Big to Fail’. *2012 Annual Report of the Federal Reserve Bank of Dallas.* Dallas, TX: Federal Reserve Bank, pp. 5-10.

Garrett, T. A. and T. L. Marsh (2002): ‘The revenue impacts of cross-border lottery shopping in the presence of spatial autocorrelation’. *Regional Science and Urban Economics*, vol. 32, pp. 501-519.

Glass, A. J. and K. Kenjegalieva (2019): ‘A spatial productivity index in the presence of efficiency spillovers: Evidence for U.S. banks, 1992-2015’. *European Journal of Operational Research*, vol. 273, pp. 1165-1179.

Glass, A. J., K. Kenjegalieva and J. Paez-Farrell (2013): ‘Productivity growth decomposition using a spatial autoregressive frontier model’. *Economics Letters*, vol. 119, pp. 291-295.

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

He, Z., I. G. Khang and A. Krishnamurthy (2010): ‘Balance sheet adjustments during the 2008 crisis’. *IMF Economic Review*, vol. 58, pp. 118-156.

Hirtle, B. (2007): ‘The impact of network size on bank branch performance’. *Journal of Banking and Finance*, vol. 31, pp. 3782-3805.

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

Kao, Y-H. and A. K. Bera (2013): ‘Spatial regression: The curious case of negative spatial dependence’. University of Illinois, Urbana-Champaign, Mimeo.

Kelejian, H. H. and I. R. Prucha (2001): ‘On the asymptotic distribution of the Moran’s I test statistic with applications’. *Journal of Econometrics*, vol. 104, pp. 219-257.

Kovner, A., J. Vickery and L. Zhou (2014): ‘Do big banks have lower operating costs?’. *Federal Reserve Bank of New York Policy Review*, vol. 20, pp. 1-27.

Kumbhakar, S. C. and C. A. K. Lovell (2000): *Stochastic Frontier Analysis.* Cambridge, UK: Cambridge University Press.

Langfield, S. and M. Pagano (2015): ‘Bank bias in Europe: Effects on systemic risk and growth’. European Central Bank Working Paper No. 1797.

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

Lozano-Vivas, A. (1997): ‘Profit efficiency for Spanish savings banks’. *European Journal of Operational Research*, vol. 98, pp. 381-394.

Lozano-Vivas, A. and F. Pasiouras (2014): ‘Bank productivity change and off-balance-sheet activities across different levels of economic development’. *Journal of Financial Services Research*, vol. 46, pp. 271-294.

Mester, L. J. (2010): ‘Scale economies in banking and financial regulatory reform’. *The Region.* Minneapolis, MN: Federal Reserve Bank, pp. 10-13.

Pesaran, M. H., T. Schuermann and S. M. Weiner (2004): ‘Modeling regional interdependencies using a global error-correcting macroeconometric model’. *Journal of Business and Economic Statistics*, vol. 22, pp. 129-162.

Qu, Xi. and L-F. Lee (2015): ‘Estimating a spatial autoregressive model with an endogenous spatial weight matrix’. *Journal of Econometrics*, vol. 184, pp. 209-232.

Sealey, C. and J. T. Lindley (1977): ‘Inputs, outputs and a theory of production and cost at depository financial institutions’. *Journal of Finance*, vol. 32, pp. 1251-1266.

Stern, G. H. and R. Feldman (2009): ‘Addressing TBTF by shrinking financial institutions: An initial assessment’. *The Region.* Minneapolis, MN: Federal Reserve Bank, pp. 8-13.

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

Wheelock, D. C. and P. W. Wilson (2009): ‘Robust nonparametric quantile estimation of efficiency and productivity change in U.S. commercial banking, 1985-2004’. *Journal of Business and Economic Statistics*, vol. 27, pp. 354-368.

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

Wheelock, D. C. and P. W. Wilson (2018): ‘The evolution of scale economies in US banking’. *Journal of Applied Econometrics*, vol. 33, pp. 16-28.

Whittle, P. (1954): ‘On stationary processes in the plane’. *Biometrika*, vol. 41, pp. 434-449.