



Transitive properties: a spatial econometric analysis of new business creation around transit

Kevin Credit

To cite this article: Kevin Credit (2019) Transitive properties: a spatial econometric analysis of new business creation around transit, *Spatial Economic Analysis*, 14:1, 26-52, DOI: [10.1080/17421772.2019.1523548](https://doi.org/10.1080/17421772.2019.1523548)

To link to this article: <https://doi.org/10.1080/17421772.2019.1523548>



Published online: 04 Oct 2018.



Submit your article to this journal [↗](#)



Article views: 721



View related articles [↗](#)



View Crossmark data [↗](#)



Citing articles: 7 View citing articles [↗](#)



Transitive properties: a spatial econometric analysis of new business creation around transit

Kevin Credit 

ABSTRACT

This paper evaluates the relationship between transit station proximity and new business creation in five US regions with varying levels of maturity in rail transit development and/or entrepreneurial ecosystems: Boston, San Jose, Austin, Cleveland and Philadelphia. It tests a variety of spatial econometric models to find the best specification and compares the results with the kinds of non-spatial models currently used in the literature. This provides a better understanding of the role of various forms of spatial dependence in the transit – new business creation relationship and shows that existing models may overstate the impact of transit on new business creation. In addition, the paper teases out differences between regions, rail modes and business types that can be usefully applied to a variety of urban contexts.

KEYWORDS

spatial models (C21), economic development (F63), urban, rural, regional, and transportation analysis (O18), planning policy (O21)


HISTORY Received 9 August 2017; in revised form 15 August 2018

INTRODUCTION

Interest in rail transit has increased in recent years, as concerns about sustainability, traffic congestion and sprawl have come to the forefront of urban research (Batty, Besussi, & Chin, 2003; Ewing & Cervero, 2010; Richardson & Bae, 2016; Squires, 2002). North American cities have invested large sums on new transit systems, and scholars have undertaken a vast research programme to examine their impact on travel behaviour, development and property values in detail (Ewing & Cervero, 2010; Freemark, 2014; Mohammad, Graham, Melo, & Anderson, 2013). The economic impact of transit systems is particularly important to understand, given their large fixed cost and the desire of policy-makers to use transit systems as economic development and urban revitalization tools.

Existing research has identified a strong but nuanced link between rail transit and increasing property values (Ewing & Cervero, 2010; Golub, Guhathakurta, & Sollapuram, 2012; Mohammad et al., 2013); however, the impact of transit on new business formation has been relatively less explored. Given the fact that transit increases accessibility, it has the potential to foster a number of benefits to businesses that locate near it: increased face-to-face contact, ‘weak’ business ties with clients, labour market accessibility, higher consumer visibility (for retail products), knowledge spillovers, and recruitment for younger workers who enjoy car-free living (Chatman & Noland, 2011;

CONTACT

(Corresponding author)  kcredit@uchicago.edu
Center for Spatial Data Science, University of Chicago, Chicago, IL, USA.

Credit, 2017). In fact, some researchers have suggested that certain kinds of knowledge and high-technology businesses may be moving toward a transit-centric location model in order to take advantage of these economic benefits (and to avoid the negative externalities of increasing traffic congestion) (Weisbrod, Duncan, & Moses, 2014).

Studies that have begun to look at the link between rail transit and new business creation thus far have focused primarily on the construction of new transit systems in cities such as Phoenix, Dallas and Portland (Chatman, Noland, & Klein, 2016; Credit, 2017). Less is known about the relationship between transit proximity and new business creation in regions with more established rail transit networks. The power of transit stations to attract new business investment well after construction is critical to understand, since transit systems represent large long-term fixed costs, and previous research on new transit construction has found that transit-related investment often peaks in the *run-up* to a new system's construction, declining afterwards as the system becomes a common part of the regional transportation network (Credit, 2017; Golub et al., 2012; Mohammad et al., 2013).

Thus, the goal of this paper is to examine the relationship between transit station proximity and new business creation in five regions with varying levels of maturity in rail transit development and/or entrepreneurial ecosystems – Boston, San Jose, Austin, Cleveland and Philadelphia – using point-resolution data. The relationship between proximity to transit – location within 0.25 or 0.5 mile of stations – and new business starts is evaluated using a suite of spatial econometric models at the census block level, with comparisons of effect made between each metro, mode (light rail, heavy rail or commuter rail), and business type (all knowledge businesses, high-technology, retail/services/food and producer services).

The results of the analysis indicate that, using a spatially responsive estimate of the expected density of new businesses per acre as the dependent variable, proximity to transit stations is significantly related to new business creation in a variety of contexts. As indicated by calculating the percentage of positive and significant rail coefficients across all regional models (out of all possible rail coefficients), new businesses in Philadelphia and Boston – established transit regions with more extensive networks, supportive land use and generally better service – show the largest association with rail transit. Retail, services and food businesses are most commonly significantly associated with transit proximity, while the commuter rail mode is most consistently associated with adjacent new business activity of any type. In addition, the use of a spatial Durbin model (compared with a non-spatial Poisson model with spatial lags of the dependent variable) changes the significance of transit proximity variables in 55% of instances, indicating that the use of non-spatial models may, in many cases, overestimate the positive impact of transit proximity on new business creation.

LITERATURE

An extensive literature, stretching back to the 1970s, has examined the economic impacts of rail transit (Cervero, 1984, 1994; Knight & Trygg, 1977). This research has generally found that property values increase with proximity to transit stations, especially if the service is good and the land-use planning around the stations is done in a way that cohesively integrates new development with the transit station (Agostini & Palmucci, 2008; Cervero, 2004; Damm, Lerman, Lerner-Lam, & Young, 1980; Golub et al., 2012; Knaap, Ding, & Hopkins, 2001; Landis, Guhathakurta, & Zhang, 1994; Weinberger, 2001; Weinstein & Clower, 2003). Generally, property value benefits are larger for commercial properties and commuter rail systems (Mohammad et al., 2013). However, a significant vein of research challenge these assertions, arguing that transit systems simply reflect or refocus the economic growth of the region (Giuliano, 2004; Schuetz, 2014; Vessali, 1996), or that transit does not play a primary causative role in fostering walkability,

dense development and supplemental economic activity (Bollinger & Ihlanfeldt, 1997; Chatman, 2013; Quinn, 2006). In addition, previous research has found evidence of a so-called ‘novelty factor’ for new transit systems – property values are often highest at the time a new system opens (or right before opening), indicating that perception may be somewhat larger than economic reality for transit (Golub et al., 2012; Mohammad et al., 2013).

As the impact of rail transit on property values has been the focus of the bulk of research on transit’s economic impacts, the relationship between transit and new business creation is comparatively understudied. Existing research has generally taken one of two approaches: cross-sectional multinomial logit analysis where the dependent variable is the type of business (or, in some cases, the type of industrial agglomeration) (Mejia-Dorantes, Paez, & Vassallo, 2012; Song, Lee, Anderson, & Lakshmanan, 2012), or time-series Poisson or negative binomial analysis where the dependent variable is the aggregated count of new businesses in a particular industry in a spatial unit (Chatman et al., 2016; Credit, 2017). In both cases, a measure of transit accessibility is used as the independent variable of interest. Given the relatively small body of the existing literature, each of the relevant studies is examined here in some detail.

Mejia-Dorantes et al. (2012) show that proximity to a new *Metrosur* (heavy-rail) transit station in the Madrid, Spain, region is a significant predictor of business activity for all industries, particularly for accommodation/food service and retail businesses. Interestingly, proximity to the regional commuter rail service decreases the probability of businesses locating nearby, due perhaps to the negative externalities of an above-ground rail line. In an analysis of regional Seoul, South Korea, Song et al. (2012) find industry-specific effects for subway accessibility – a village’s concentration in the construction, transportation, sales, accommodation/food service, finance, real estate or other service industries are all significantly (positively) related to subway accessibility, either directly in the village itself or through a neighbouring village.

Similarly, Credit (2017) and Chatman et al. (2016) find that proximity to transit stations is generally a strong predictor of new business activity in the retail, service, information and finance/insurance industries. Both papers use time-series data to evaluate the post-construction impact of light-rail systems in US regions that only recently built new rail-transit systems: Phoenix (Credit, 2017), Portland and Dallas (Chatman et al., 2016). In general, these studies also find that the impact of transit on new business creation declines as distance from the stations increases (0.25–1 mile). By using a quasi-experimental study design, Credit (2017) also finds that the impact of transit proximity on new business creation is highest at the time of the system’s opening, declining steadily as time from opening increases. This corroborates previous property value research that shows a similar ‘novelty factor’ for station-area property values (Golub et al., 2012; Mohammad et al., 2013).

From a methodological standpoint, advances have been made in recent years in the specification and estimation of spatial econometric models. The spatial models most commonly found in the literature – the spatial autoregressive lag (SAR) and the spatial error model (SEM) – are each designed to isolate a specific form of spatial interaction, either of which, if present, violates the independence assumption of ordinary least squares (OLS) regression (Anselin & Rey, 2014). This is accomplished mathematically by inserting a spatial weights matrix directly into the estimation of the model (Elhorst, 2014, 2017). The SAR model contains an endogenous interaction effect (spatial dependence in the dependent variable), while the SEM accounts for spatial interaction contained in the error term (Anselin & Rey, 2014). While these models improve on non-spatial regression methods and generally do not pose theoretical issues with econometric estimation, recent work – aided by the development of new estimators – has argued that these models are too simplistic (Elhorst, 2014, 2017) and are only appropriate in cases where the researcher has no uncertainty about model specification (LeSage, 2014). Due to this, the focus in spatial econometrics has begun to shift to developing and using models that account for multiple types of spatial interaction: the spatial lag of X (SLX) (containing only *exogenous* interaction effects, i.e.,

dependence from the neighbouring independent variables that affects the dependent variable) (LeSage & Pace, 2009), spatial Durbin (SDM) (containing both endogenous and exogenous interaction effects) (Anselin, 1988; LeSage & Pace, 2009), and spatial Durbin error models (SDEM) (containing endogenous, exogenous, and error effects) (Elhorst, 2014). While there remains a debate regarding the presence of identification issues in these more complex models (Anselin & Rey, 2014), if correctly specified, they have the potential to provide a richer estimation of spatial interaction effects than the older SAR and SEM approaches.

Given the state of existing research, several significant questions remain unanswered. First, the presence of a novelty factor – and whether this influences the relationship between new business creation and transit stations in *established* transit regions – needs to be explored further. Second, while previous research has examined the industry-specific impacts of transit proximity, these analyses have generally been done at fairly aggregated industrial scales. Given the theoretical links between transit accessibility and agglomeration benefits such as knowledge spillovers (through informal interactions), face-to-face contact and the construction of social trust, and the recruitment and marketing of businesses to a younger demographic (Chatman & Noland, 2011; Credit, 2017), it has been hypothesized that knowledge-intensive businesses, such as high-technology and producer services firms, might benefit particularly from proximity to transit (Weisbrod et al., 2014). Despite the importance of these sectors to regional growth (Chapple, Markusen, Schrock, Yamamoto, & Yu, 2004; DeVol & Wong, 1999; O’Hallachain & Reid, 1991), no study has yet looked at these fine-grained industrial categories, or compared their effects with more traditional business types, such as retail and services. Similarly, the explicit differences in effect between rail modes – light, heavy and commuter – have not been explored in depth. Finally, while existing studies have included some spatial variables in their modelling approaches to control for the effects of spatial autocorrelation, none of the existing papers has used an explicitly spatial econometric approach. While this is understandable given the fact that the models employed so far have been non-linear (thus incompatible with some assumptions of OLS regression inherent for common spatial econometric models), the use of spatial models to explore the link between new business creation and transit proximity remains an important gap in the literature.

DATA AND STUDY AREAS

Data

Three primary types of data are used in this analysis: information on transit systems, individual business data and socio-demographic covariates provided by the census. Table 1 describes the data and each of their sources in detail. The data on transit station locations by mode come primarily from each region’s respective transportation authorities; in some cases it was provided as a geographic shapefile, and in others it had to be digitized manually (with corroboration from Open Street Map and other sources). Classification of the modes is based primarily on definitions provided by the American Public Transit Agency (APTA) (1994): light-rail trains or trolleys are electrically powered and often run in the street right-of-way; heavy-rail transit is generally separated from traffic and is designed for larger passenger capacities than light rail; and commuter rail is also grade separated but designed specifically to link a central business district with outlying residential areas. In order to analyze the impact of transit proximity on new business creation, 0.25- and 0.5-mile buffers were calculated around each digitized transit station. For the purposes of this analysis, a census block is considered ‘within’ a given buffer if its centroid falls within the buffer.

The business data used in this paper come from the National Establishment Time Series (NETS), which provides information on industrial classification (by North American Industry Classification System (NAICS) code), location and year of opening (among other features). The underlying business data for NETS is furnished by the Dun and Bradstreet database and geocoded by Walls and Associates, providing a near census of business activity in a metropolitan area

Table 1. Data sources and descriptions.

Category	Region/ scale	Description ^a		Date	Source
		Type	Description		
Transit	Austin	Commuter rail	Capital MetroRail	2010	Digitized: https://www.capmetro.org/schedmap/?svc=2&f1=550&s=0&d=N
	Boston	Commuter rail	Massachusetts Bay Transportation Authority (MBTA) Commuter Rail	2014	Downloaded: http://www.mass.gov/anf/research-and-tech/it-serv-and-support/application-serv/office-of-geographic-information-massgis/datalayers/mbta.html
		Heavy rail	MBTA Subway (Red, Orange and Blue Lines)	2014	
		Light rail	MBTA Trolley (Green Line)	2014	
	Cleveland	Heavy rail	Regional Transit Authority (RTA) Rapid Transit (Red Line)	2016	Digitized: http://www.riderta.com/sites/default/files/gtfs/latest/google_transit.zip
		Light rail	RTA Rapid Transit (Blue, Green and Waterfront Lines)	2016	
	Philadelphia	Commuter rail	Southeastern Pennsylvania Transportation Authority (SEPTA) Regional Rail	2009	Downloaded: https://www.arcgis.com/home/item.html?id=1cb6bfb987c44859a2fffa7384cc5cd2
		Heavy rail	SEPTA High Speed Rail	2009	
			Port Authority Transit Corporation (PATCO) Speedline	2015	Downloaded: https://njgin.state.nj.us/NJ_NJGINExplorer/DataDownloads.jsp
		Light rail	SEPTA Trolley (10, 11, 13, 15, 34, 36, 101 and 102 Lines)	2009	Downloaded: https://www.arcgis.com/home/item.html?id=1cb6bfb987c44859a2fffa7384cc5cd2
San Jose	Commuter rail	Caltrain and Altamont Corridor Express Commuter Rail	2013	Downloaded: http://www.dot.ca.gov/hq/tsip/gis/datalibrary/Metadata/RR_Commuter_13.html	
	Light rail	Valley Transportation Authority (VTA) Light Rail	2015	Downloaded: https://data.vta.org/Transit-Operations/Stops-January-2015/iy3q-7kq5	

Business	All knowledge	NAICS codes 51, 52, 54 and 55	2011	National Establishment Time Series (NETS)
	High-tech	NAICS codes 3254, 3341, 3342, 3344, 3345, 3364, 5112, 5161, 5179, 5181, 5182, 5413, 5415 and 5417	2011	
	Producer services	NAICS codes 5411, 5412, 5414, 5416, 5418, 5419 and 5511	2011	
	Retail/services	NAICS codes 4431, 4451, 4452, 4453, 4461, 4481, 4482, 4483, 4511, 4512, 4522, 4531, 4532, 4533, 4539, 8114, 8121, 8123, 8129, 8133, 8134, 8139, 7211, 7213, 7223, 7224 and 7225	2011	
Socio-demographic	Block level	Total population; racial diversity of block (computed using the Herfindahl index); size of block (acres)	2010	Decennial census
	Tract level	Percentage bachelor's degree attainment or higher; percentage of the population aged 29 years or younger; average number of vehicles per household	2008–12	American Community Survey

Notes: ^aFor transit data, this year specifies the date the file used in the analysis was created; however, transit stations used in the analysis existed in 2010. NAICS, North American Industry Classification System.

given year (Walls & Associates, 2012). For this paper, businesses that started in 2011 in four industries of interest – all knowledge, high-tech,¹ producer services and retail/personal services – were delineated and spatially joined (using ArcMap v.10.3) to the census blocks in which they are located. Aggregated counts of new businesses² in each industry serve as the dependent variable. Year 2011 new businesses were selected in order to provide at least a one-year lag between neighbourhood characteristics (the census covariates described below) and the opening of a new business, since the decision to open a business in a given location is most likely based on the observed characteristics of the neighbourhood at least one year before the business actually opens. Because NETS geocodes a business' location to its *last* place of establishment, no business points that relocated were included in the analysis.

Socio-demographic data for 2010 were collected at two scales: the census block and the census tract. Basic demographic information – including racial classification and total population – is provided at the block level from the decennial census. A Herfindahl index was used to create a measure of racial diversity (where values closer to 0 indicate higher levels of diversity) at this scale. Unfortunately, the remaining socioeconomic characteristics of interest are only provided at the tract level by the 2008–2012 American Community Survey (ACS), so these variables were joined to their nested census blocks.

Study areas

In order to ascertain the relationship between transit proximity and new high-technology businesses in mature transit regions, five study areas with fixed-rail transit systems operating in 2010 were selected to cover a full range of characteristics: regions with established entrepreneurial ecosystems but a relative lack of transit prominence, such as San Jose, California, and Austin, Texas; regions with established transit systems but little historical entrepreneurial activity, such as Cleveland, Ohio, and Philadelphia, Pennsylvania; and a region with *both* prominent entrepreneurial activity and a mature transit system with a supportive built environment, Boston, Massachusetts (Chapple et al., 2004; Richman, 2015; Saxenian, 1994). This choice of regions also allows for considerable modal variation: four regions (San Jose, Cleveland, Philadelphia and Boston) have light-rail systems, three (Cleveland, Philadelphia and Boston) have heavy-rail systems, and four (San Jose, Austin, Philadelphia and Boston) have commuter rail systems.

As for the technical delineation of these regions, the census blocks for all counties in each Metropolitan Statistical Area (MSA) containing rail transit stations were used. This allows the paper to employ a regional approach (rather than biasing the sample towards transit use by selecting, for example, only neighbourhoods around transit stations) while limiting some suburban/exurban bias that could crop up if the entire MSA definition were used, given that many of the transit systems examined here extend only into the inner-ring suburbs of their respective regions. It makes little sense to include areas that are truly transit inaccessible – such as exurban or rural portions of an MSA – in this analysis, since businesses in these areas are highly unlikely to consider transit as a part of their location calculus.

METHODS

The model specification for this paper involves some interesting trade-offs – using the individual business points themselves preserves the finest grain of spatial accuracy in relation to transit stations, but unfortunately it does not answer the research question at hand. If the unit of analysis were the businesses themselves, the paper would involve studying how characteristics of businesses influence location – for any given *individual* business – close to transit; rather, this paper is interested in what (if anything) influences *higher numbers of businesses* to be located near transit stations because this focus on neighbourhood and regional determinants of new business creation is important for helping to guide local economic development and planning efforts (Mack & Credit, 2016;

Malecki, 1984; Renski, 2008). To do this, a unit of aggregation must be selected; however, if the aggregation unit is too large (e.g., census tracts), location within 0.25 or 0.5 mile of a transit station loses its descriptive power. That is why this paper uses the smallest possible aggregation unit, census blocks, in order to maintain as much spatial variation as possible in measuring proximity to transit. Covariate characteristics that can be measured at the block level are included at the block level (such as race, population and median age), while other measures only available at the tract level are duplicated for nested blocks.

Of course, the choice of such a small aggregation unit comes with its own challenges. Given the very fine spatial scale of blocks, in a given year (2011) there are a vast number of zero new business counts. This presents a problem for statistical modelling similar to the ‘small numbers problem’ common in the spatial epidemiology literature – for very rare events (e.g., cancer deaths), the spatial distribution observed in any one timeframe may not reflect the true underlying probability of the event occurring (Lawson, Banerjee, Haining, & Ugarte, 2016). This has the potential to bias regression models constructed based on the observed data, since the very rare observed events will drive, perhaps unfairly, the model results. In these cases, smoothing functions are often employed in order to approximate the ‘true’ underlying probability of an event occurring at any given location by adjusting expected event counts in some way (Clayton & Kaldor, 1987; Kafadar, 1996; Lawson et al., 2016).

The simplest smoothing function is simply to calculate the raw rate:

$$z_i = \frac{x_i}{y_i} \quad (1)$$

where x_i is the count of events of interest at location i (in this case, the count of new high-technology businesses); and y_i is the exposure variable (in this case, size of the block in acres). This operation simply standardizes the raw count of events by some underlying exposure rate. Of course, the underlying probability of an event occurring at i may very likely be influenced by its neighbours. In this case, a ‘spatial rate’³ can be calculated, which uses information from neighbouring observations to create a smoothed rate (Kafadar, 1996):

$$\tilde{z}_i = \frac{ave_R\{x_i\}}{ave_R\{y_i\}} \quad (2)$$

where *ave* is an averaging function; and R is the neighbourhood of observations around either x_i or y_i . In this case, *ave* is simply the average for the queen-contiguous neighbours of a given block, computed using a spatial weights matrix. An even more detailed smoothing procedure is the spatial empirical Bayes (SEB) estimator, which ‘shrinks’ the expected value in a given location towards the mean of its neighbourhood’s rate – the ‘prior’ distribution in this case – based on the size of the variance of the raw rate at a given location, thus limiting the impact of observations with extremely high variance (Anselin, Kim, & Syabri, 2004; Clayton & Kaldor, 1987):

$$\tilde{\pi}_i = w_i z_i + (1 - w_i) \theta, \quad (3)$$

where:

$$w_i = \frac{\emptyset}{\emptyset + (\theta/y_i)}. \quad (4)$$

where θ is the mean and \emptyset is the variance of the prior distribution (in this case, the queen-contiguous neighbours of location i). As Anselin et al. (2004) point out, when the exposure variable is large, θ/y_i nears 0, which means that w_i moves toward 1, pushing nearly all the weight to the raw rate (z_i). Thus, the SEB estimate uses both the mean and the variance of the spatial neighbourhood to weight a given observation, producing a more-informative expected rate estimate. As

For each business type (knowledge, high technology, retail/services, and producer services):

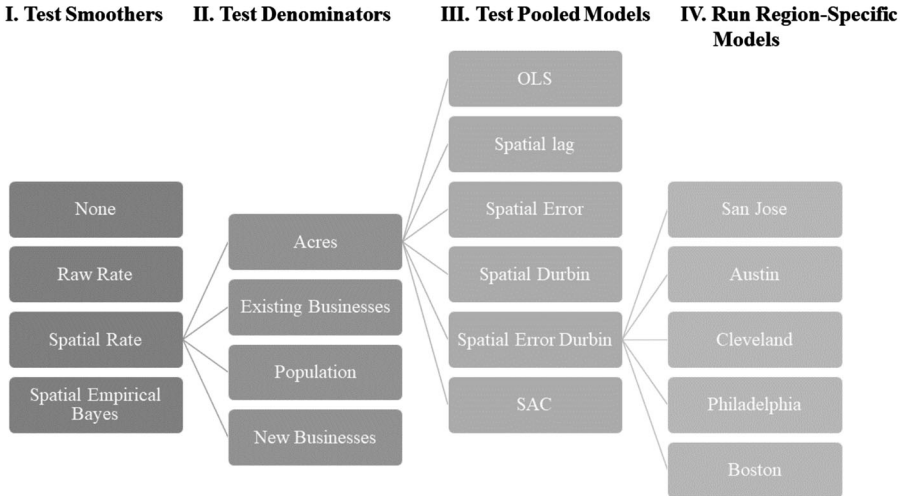


Figure 1. Analysis framework for testing the rate smoothing and spatial econometric approaches.

described in the section ‘Exposure variables for new business creation’ below, each of these smoothing techniques was calculated for each business type, producing expected rates of new business creation, which were then used as the dependent variable in a series of linear models comparing the relative performance of each smoothing technique in order to find the best method for the final model specification. Figure 1 shows this process diagrammatically: first, different spatial and non-spatial smoothing techniques are tested for each exposure variable. The diagnostics of OLS models run using the best-performing smoothers for each exposure variable are then compared to find the most consistent form of the dependent variable for each business type. These dependent variables are used in a suite of pooled spatial econometric models to find the best-fit spatial model form; once the final model specification is selected, region-specific models are run and compared in order to compare and contrast results across regions.

Modelling spatial count data

In standard econometric practice, generalized linear models (GLM) such as Poisson or negative binomial models are chosen to estimate count data, given the fact that the dependent variable is modelled as the expected outcome of a Poisson distribution, which is non-negative and has a mean equal to its variance (Faraway, 2006; Hilbe, 2014).⁴

However, the use of models that directly incorporate a spatial dependence structure into the model specification in the form of a spatial weights matrix is suggested where data are spatially autocorrelated, as is the case with new businesses (Anselin & Rey, 2014). There are three main approaches to estimating GLM models that take into account underlying spatial dependence in the data. The simplest is to include spatially lagged versions of the dependent and/or independent variables in a non-spatial Poisson or negative binomial model in order to control for spatial dependence. Since these models do not directly incorporate spatial dependence into the covariance structure of the data, they are the least favoured.

Spatial generalized linear mixed models (GLMM), on the other hand, add a random effects parameter to the linear predictor of the Poisson model, which is assumed to be normally distributed and whose covariance is calculated using a spatial weight matrix (Liu, Davidson, & Apanasovich, 2007). There is considerable complexity, however, in estimating the model: maximum likelihood estimation results in an *n*-dimensional integral (where *n* is the number of observations)

that cannot be reduced due to the spatial correlation between the observations, which is difficult to integrate with any large n . Bayesian hierarchical spatial models that employ Markov chain Monte Carlo methods to simulate samples of the posterior distribution of the regression parameters and the covariance structure of the Gaussian random field are more widely used, but the impact of data composition on the model, long computation times and the fact that basic statistical properties of these models are still not fully understood continue to be issues for implementation (Christensen, Roberts, & Sköld, 2006; De Oliveira, 2013; Diggle & Ribeiro, 2007; Diggle, Tawn, & Moyeed, 1998; Jing & De Oliveira, 2015). Other composite likelihood approaches reduce the computational effort for estimation (Liu et al., 2007), but even so, spatial GLMM approaches remain challenging to use for many spatial econometric applications.

The third approach is spatial filtering, which uses the eigenvectors of the data's spatial relationships (via a spatial weights matrix) to remove systematically residual dependence in the model (Tiefelsdorf & Griffith, 2007; Wang, Kockelman, & Wang, 2013). Since this approach produces a set of spatial eigenvectors that can then be added to any model type – including Poisson or negative binomial models – this approach can be used to eliminate spatial dependence in models of count data. However, since the number of eigenvectors calculated equals the number of observations – from which the top vectors are identified to include in the final model – the spatial filtering approach can involve lengthy computation times (Wang et al., 2013).

Given the fact that some computational issues persist with each of these approaches to modelling spatial count data, this paper presents an alternative method for analyzing business count data in a traditional spatial econometric framework⁵ by using rate smoothing approaches (mentioned above) to help alleviate the problems of non-normality and small numbers. A \log_{10} -transformation of the expected counts provided by a 'best-fit' smoothing technique then provides a normally distributed variable that can be used for standard spatial econometric model estimation.

In order to choose a suitable smoothing function, however, an exposure variable for new business creation must be chosen. While various theoretical arguments could be made to support the use of area, population, the number of existing businesses or the number of all new businesses in a block as a suitable background rate – for instance, businesses driven by consumer demand might logically employ population as an exposure measure – this paper is interested in empirically testing these measures in order to see which performs best in a set modelling framework.

Exposure variables for new business creation

In order to evaluate empirically the logical choices of background exposure variables to construct a suitable dependent variable for use in a spatial regression, the best-performing method from each of the four possible exposure variables is chosen – along with the log-transformed raw count – to construct four regressions in order to evaluate the sensitivity of the results to changes in the exposure variable. The full range of possible exposure variables and rate smoothing techniques are shown in Table 2, along with some of their basic characteristics.

For each business type, the smoothing technique⁶ with the greatest number of non-zero observations and/or transformed normality for each of the four possible exposure variables are chosen for testing. A \log_{10} -transformed version of these four variables then become the dependent variables in four OLS regressions, the comparative diagnostics of which are shown in Table 3. Each model is compared based on a range of diagnostics, including adjusted R^2 and residual standard error (for overall model fit), multicollinearity condition, Breusch–Pagan and Koenker–Bassett tests (for heteroskedasticity), and comparison of the signs and significances of the coefficients with a Poisson regression run with the same specification and the exposure variable added as an offset.

As Table 3 indicates, spatial rate smoothing using area as the exposure variable is shown to be the best-performing method for knowledge, retail/services/food and producer services businesses. For high-technology businesses, the SEB smoothing technique with area as the exposure variable

Table 2. Rate-smoothing methods and exposure variables considered for testing.

Business type	Method	Exposure variable	Non-zero observations	Distribution (after log(10) transformation)	
Knowledge (NAICS 51–52 and 53–54)	Raw count	None	14,687	Right skewed	
	Raw rate	Acres	14,687	Normal	
	Spatial rate	Acres	95,889	Normal	
	SEB	Acres	95,889	Slightly left skewed	
	Raw rate	All existing businesses in 2011	14,687	Normal	
	Spatial rate	All existing businesses in 2011	95,889	Slightly right skewed	
	SEB	All existing businesses in 2011	9171	Slightly left skewed	
	Raw rate	2010 population	12,712	Normal	
	Spatial rate	2010 population	93,792	Right skewed	
	SEB	2010 population	14,127	Normal	
	Raw rate	All new businesses in 2011	14,687	Heavily left skewed	
	Spatial rate	All new businesses in 2011	95,889	Left skewed	
	SEB	All new businesses in 2011	196	Left skewed	
	High-technology	Raw count	None	2962	Right skewed
		Raw rate	Acres	2962	Slightly left skewed
Spatial rate		Acres	26,721	Normal	
SEB		Acres	26,725	Normal	
Raw rate		All existing businesses in 2011	2962	Normal	
Spatial rate		All existing businesses in 2011	26,721	Normal	
SEB		All existing businesses in 2011	2290	Normal	
Raw rate		2010 population	2381	Normal	
Spatial rate		2010 population	25,735	Slightly right skewed	
SEB		2010 population	3364	Normal	
Raw rate		All new businesses in 2011	2962	Heavily left skewed	
Spatial rate		All new businesses in 2011	26,725	Left skewed	
SEB		All new businesses in 2011	79	Left skewed	
Retail, services and food		Raw count	None	8044	Heavily right skewed
		Raw rate	Acres	8044	Slightly left skewed
	Spatial rate	Acres	62,240	Normal	
	SEB	Acres	62,240	Left skewed	
	Raw rate	All existing businesses in 2011	8044	Slightly left skewed	
	Spatial rate	All existing businesses in 2011	62,240	Normal	
	SEB	All existing businesses in 2011	5939	Slightly left skewed	
	Raw rate	2010 population	6727	Slightly right skewed	
	Spatial rate	2010 population	60,755	Right skewed	
	SEB	2010 population	8656	Normal	

(Continued)

Table 2. Continued.

Business type	Method	Exposure variable	Non-zero observations	Distribution (after log(10) transformation)
Producer services	Raw rate	All new businesses in 2011	8044	Heavily left skewed
	Spatial rate	All new businesses in 2011	62,240	Left skewed
	SEB	All new businesses in 2011	135	Slightly left skewed
	Raw count	None	8936	Heavily right skewed
	Raw rate	Acres	8936	Slightly left skewed
	Spatial rate	Acres	67,139	Normal
	SEB	Acres	67,139	Normal
	Raw rate	All existing businesses in 2011	8936	Normal
	Spatial rate	All existing businesses in 2011	67,139	Slightly right skewed
	SEB	All existing businesses in 2011	6695	Slightly left skewed
	Raw rate	2010 population	7847	Right skewed
	Spatial rate	2010 population	65,850	Right skewed
	SEB	2010 population	9817	Normal
	Raw rate	All new businesses in 2011	8936	Heavily left skewed
Spatial rate	All new businesses in 2011	67,139	Left skewed	
SEB	All new businesses in 2011	166	Normal	

Notes: The four ‘best’ method/variable combinations selected for comparison (for each of the four dependent variables of interest) are marked in bold.

$N = 227,140$ total observations in the data set.

SEB, spatial empirical Bayes.

performed best, perhaps due to the very small number of observations for that business type. Beyond the quantitative diagnostic advantage of the models that use area as the exposure variable, the area-smoothed dependent variables in these models also simply make greater qualitative sense: the expected value of new business density is an easier concept to pin down in reality than ‘the expected proportion of new businesses to population’ in a block. All this suggests that area provides the most stable exposure rate from which to measure new businesses, which has useful implications for future spatial econometric research on new business creation.

Spatial regression specification

With the proper exposure variable and smoothing technique for each type of new business chosen, the specification for the full spatial regression models can be determined. Given the lack of prior theoretical knowledge about the correct model and spatial weights matrix combination to choose, this paper follows LeSage’s (2014, 2015) method for computing Bayesian posterior model probabilities to compare three types of pooled models (containing observations for all five regions) – the SLX, SDM and SDEM – and a range of spatial weights matrices, from three to 20 nearest-neighbours⁷ for each dependent variable of interest. While alternative methods for solving the model comparison problem have been proposed (Elhorst et al., 2016), including a non-Bayesian method from Gerkman and Ahlgren (2014), LeSage’s approach is (to this point) the most theoretically

Table 3. Comparison of the model results for the four ‘best’ method/variable combinations for each of the four dependent variables of interest.

Business type	Smoothing rates	N	Adjusted R ²	Residual standard error	Multicollinearity condition	Breusch–Pagan	Koenker–Bassett	Variables with changing significance ^a	Unexpected sign for significant variables	Overall Score ^b	Other factors		Model selection
											Pros	Cons	
Knowledge (NAICS 51–52 and 54–55)	Spatial rate – Acres	95,889	0.2	0.555	12.296	7230.916	6503.683	None	PCTBACH, POP	11	Normality, intuitive smoothed denominator, large number of observations	None	***
	Spatial rate – Existing businesses	95,889	0.045	0.283	11.983	1492.607	1152.488	LR25, HR25, CR5, VEHPERHH, NS_ALL	CR25, HR5, POP	9	Large number of observations	Slightly right skewed	
	Raw rate – Population	12,712	0.154	0.509	15.165	8762.823	3336.833	LR25, PCTBACH, VEHPERHH, ACRES	LR5, EXYR11	7	Normality, intuitive rate transformation	Low number of observations	
High-technology	Spatial rate – New businesses	95,889	0.04	0.253	12.208	1369.914	1488.488	HR25, CR25, RACE_DIV, VEHPERHH	LR25, LR5, HR5, EXYR11	9	Normality, large number of observations	Left skewed	
	Spatial empirical Bayes – Acres	26,725	0.211	0.589	12.755	870.078	521.456	POP	PCTBACH, EXYR11	12	Normality, intuitive smoothed denominator	Complex rate transformation	***
	Spatial rate – Existing businesses	26,721	0.108	0.334	12.496	2511.156	1703.268	CR25, LR5, CR5, VEHPERHH, ACRES, POP	LR25, HR5, NS_ALL	9	Normality	None	
	Spatial rate – Population	25,735	0.121	0.498	12.437	3947.822	1050.356	LR25, HR5, CR5, ACRES	LR5, RACE_DIV, VEHPERHH, EXYR11	7	None	Slightly right skewed	
	Spatial rate – New businesses	26,725	0.13	0.338	12.664	5511.761	4567.508	HR5, CR5, PCTBACH, ACRES, POP, EXYR11	LR25, LR5, RACE_DIV, VEHPERHH	6	None	Left skewed	

Retail, services, and food	Spatial rate – Acres	62,240	0.22	0.541	12.103	2463.253	2277.16	POP	PCTBACH, EXYR11	11	Normality, intuitive smoothed denominator, large number of observations	None	***
	Spatial rate – Existing businesses	62,240	0.06	0.306	11.785	5096.93	3803.175	LR25, HR25, CR25, LR5, HR5, PCTBACH, VEHPERHH, POP	NS_ALL	8	Normality, large number of observations	None	
Producer services	Spatial empirical Bayes – Population	8656	0.1	0.37	14.675	2396.023	670.819	HR25, ACRES	LR25, HR5, EXYR11	6	Normality	Complex rate transformation, low number of observations	
	Spatial rate – New businesses	62,240	0.091	0.296	11.998	6787.085	6784.622	LR25, HR25, LR5, HR5, VEHPERHH, ACRES, POP	RACE_DIV	11	Large number of observations	Left skewed	
	Spatial rate – Acres	67,139	0.21	0.539	12.308	5076.311	4608.52	EXYR11	PCTBACH, POP	11	Normality, intuitive smoothed denominator, large number of observations	None	***
	Spatial rate – Existing businesses	67,139	0.059	0.298	11.992	2274.872	1682.242	HR25, CR25, HR5	LR25, LR5, CR5, RACE_DIV, PCTBACH, VEHPERHH, POP, NS_ALL	9	Large number of observations	Slightly right skewed	
	Spatial empirical Bayes – Population	9817	0.122	0.366	15.254	1848.597	670.816	HR25, HR5, CR5, VEHPERHH, ACRES	LR25, LR5, EXYR11	8	Normality	Low number of observations, complex rate transformation	
	Spatial rate – New businesses	67,139	0.08	0.295	12.198	6057.737	6151.06	CR25, HR5, CR5, VEHPERHH	LR25, HR25, LR5, PCTBACH, POP	8	Large number of observations	Left skewed	

Notes: ^aBased on a comparison with non-spatial Poisson regression run in R using all observations and the corresponding background population measure (i.e., acres, existing businesses, population, new businesses) as the offset variable. Variables chosen for regression are based on correlation analysis using the full data set.

^bScored based on the following scale: 3 points for best value for a given diagnostic; 2 points for second; 1 point for third; and 0 for worst. Cells are shaded accordingly.

***The models marked by the asterisks were the best-fit (chosen) models for each dependent variable of interest.

direct and easy to compute, and has been applied recently to dynamic spatial panel models of population growth (da Silva, Elhorst, & Neto, 2017) and government spending (Rios, Pascual, & Cabases, 2017).

To do this, posterior model probabilities comparing all 54 possible model specifications (18 possible weights matrices for the SLX, SDM and SDEM) for each dependent variable are calculated. The basic logic of this approach is to use Bayes' theorem to calculate the *posterior* probability that, for instance, M_1 is the proper model to choose given that the set of data y is true (LeSage, 2014):

$$p(M_1|y) = \frac{p(y|M_1)}{p(y|M_1) + p(y|M_2)} \cdot \frac{p(M_1)}{p(M_2)} \quad (5)$$

where $p(y|M_1)$ is the marginal likelihood for model M_1 (calculated by integrating over the vector of parameters obtained in M_1); and $p(M_1)$ is its prior probability. The marginal likelihood values – and thus posterior probabilities – for each model sum to 1.

The results of the posterior probability calculation for each weights matrix/model specification for each dependent variable are shown in Table 4. For every dependent variable, the SDM demonstrates the highest posterior probability, and thus is used as the specification for the final pooled and regional regressions used in the paper. In terms of spatial weights matrix selection, each dependent variable model shows slightly different results: for knowledge businesses, an eight nearest-neighbour row standardized weights matrix shows the highest probability, while a 13 nearest-neighbour matrix is preferred for high-technology businesses and a seven nearest-neighbour matrix is selected for both retail and producer services.

From a theoretical perspective, the use of a global spillover spatial lag-type model also makes sense, as new knowledge and high-technology businesses are expected to exhibit spatial dependence based on spillover activity, due to knowledge spillovers, information exchange and other agglomeration factors (Aharonson, Baum, & Feldman, 2007; Aharonson, Stettner, Amburgey, Ellis, & Drori, 2013; Gilbert, McDougall, & Audretsch, 2008).

The specification for the regional spatial Durbin models⁸ used in this paper is:

$$Y = \delta WY + X\beta + WX\theta + \varepsilon, \quad (6)$$

where Y is a vector of observations of the expected density of new businesses per acre in a given region; WY is the spatially lagged dependent variable (based on a seven nearest-neighbour row normalized spatial weights matrix, W); δ is the spatial autoregressive parameter; X is a matrix of exogenous covariates (including dummy variables for location within 0.25 mile and from 0.25–0.5 mile of a transit stations of particular modes); WX is vector of spatial lags of the independent variables; θ is the vector of regression coefficient for these lagged independent variables; and ε is the error term (Elhorst, 2014). The basic idea of the spatial Durbin model is that, by including lags of the independent (as well as the dependent) variables, the model can more fully account for residual spatial autocorrelation while also providing estimates of the influence that neighbouring values of the independent variables have on the expected density of new business starts. The independent variables included in the final specification are station proximity, racial diversity, per cent bachelor's degree attainment or higher, vehicles per household, population, total number of existing businesses and total number of all new businesses

It is also important to note that in spatial lag-type models it is necessary to apply scalar summary measures to the regression output, that is, to calculate the average *direct* and *indirect* effects to obtain a value of *total* partial effects comparable with a non-spatial regression coefficient (Elhorst, 2014; LeSage & Pace, 2009). This is necessary in spatial lag-type models because a change in the value of an independent variable in a given unit changes the value both of (1) that a unit's own dependent variable (the direct effect), as well as (2) the value of neighbouring units' dependent

Table 4. Results of the Bayesian posterior model and weights matrix probability comparison (LeSage, 2014, 2015).

#NN	Knowledge			High-technology			Retail			Producer services		
	SLX	SDM	SDEM	SLX	SDM	SDEM	SLX	SDM	SDEM	SLX	SDM	SDEM
3	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
4	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
5	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
6	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
7	0.00%	0.36%	0.00%	0.00%	0.00%	0.00%	0.00%	100.00%	0.00%	0.00%	100.00%	0.00%
8	0.00%	99.64%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
9	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
10	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
11	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
12	0.00%	0.00%	0.00%	0.00%	0.01%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
13	0.00%	0.00%	0.00%	0.00%	98.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
14	0.00%	0.00%	0.00%	0.00%	0.03%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
15	0.00%	0.00%	0.00%	0.00%	1.96%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
16	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
17	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
18	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
19	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
20	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

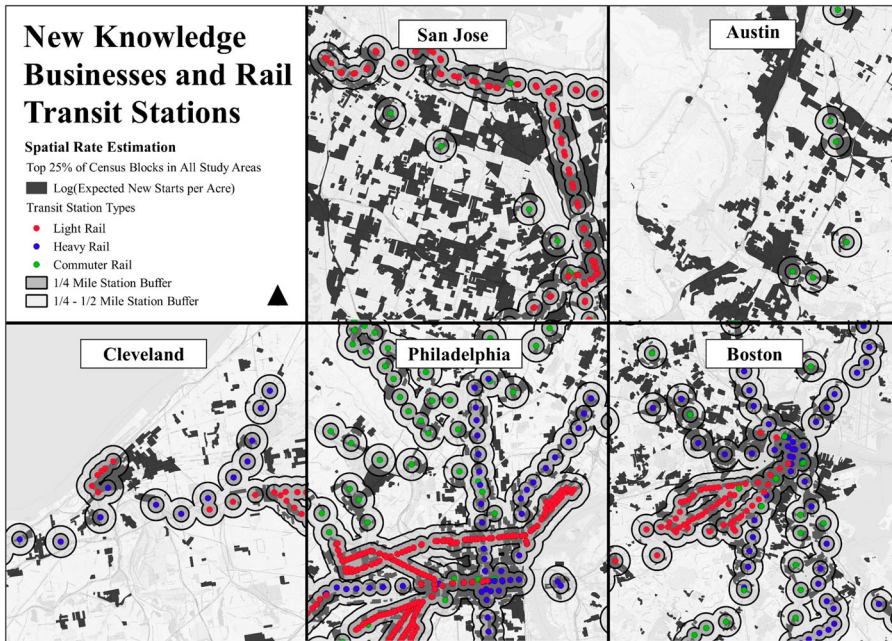


Figure 2. Census blocks in the top 25% for $\log_{10}(\text{expected new knowledge business density})$ and rail transit buffers for all modes and regions.

variables (the indirect or spillover effect). A standard regression coefficient in a non-spatial model necessarily contains both of these effects because there is no accounting for spatial interaction. These effects were calculated for each regional model using the ‘impacts’ function in the ‘spdep’ package in R.

RESULTS

Descriptive statistics and spatial patterns

Appendix A in the supplemental data online shows the combined descriptive statistics for the variables used in the spatial models for all regions. In general, it illustrates the low number of average new business starts for all types, as well as the number of blocks located within walking distance to transit stations. On average, only around 5–6% of blocks in the model are within 0.5 mile of any kind of rail transit station; commuter rail stations have a bit more reach, with 0.5-mile buffers covering over 9% of selected blocks. The average block in the data set has a fairly high educational attainment – 36% bachelor’s degree attainment or higher – and a generally average level of vehicle ownership (1.74 vehicles per household).

Figure 2 shows the broad pattern for expected new knowledge business density and transit station location in each of the five regions of interest at a similar spatial scale. While it is difficult to make conclusive statements about spatial patterns from looking at Figure 2, a few things stand out: (1) Philadelphia’s and Boston’s transit systems appear to show much more extensive coverage of blocks with high levels of new knowledge start density (in part due to the larger number of stations in those cities); (2) the pattern of top knowledge blocks and transit stations are seemingly mismatched in Austin and Cleveland; and (3) San Jose appears to have a very large number of the top knowledge blocks (as might be expected), but these also appear to be generally mismatched with the light-rail system in particular, which runs along the periphery of the new business-heavy portions of the valley.

Regression results

Table 5 shows the summarized significant, transit-specific results of the pooled and regional spatial Durbin models for each business type of interest; for detailed regression results for each of the 24 individual models, see Appendix B in the supplemental data online. Given the large number of models to compare and the relative heterogeneity of results across regions, the approach in this paper is to use the regional model results to find the percentage of positive/significant total effects coefficients⁹ for each category of interest (out of all possible positive/significant results). These percentages, as indicators of consistent association, are then used to compare across the categories of interest for the study: business type, mode and region. In addition to this analysis of the regional models, the pooled model results are also used to provide indications of overall trends by business type and mode.

In terms of business type, both the pooled and regional models indicate that new retail, food and services businesses are the most consistently associated with all types of rail transit stations, with 45% positive/significant coefficients across the regional models (and positive/significant results for every transit variable in the pooled model). All knowledge businesses are the second-most consistently associated with rail transit (at 41%), followed by high-technology and producer services businesses (27% each). While regional heterogeneity is evident in these relationships, this finding generally supports the idea that retail and services businesses rely heavily on visibility and pass-by traffic, while also providing evidence that knowledge businesses benefit from transit proximity in a variety of ways (Chatman et al., 2016; Credit, 2017).

As for modal variations, commuter rail is most consistently associated with transit proximity in both the pooled and regional models, corroborating previous research on property values that finds larger gains accrue to properties near commuter rail stations than other modes (Mohammad et al., 2013). Perhaps the generally larger regional mobility offered by commuter rail rather than heavy or light rail accounts for this result; it is also possible that, by their nature as ‘commuter’ transit, these rail stations represent generally larger investments (in terms of size and design) and/or tend to be located in more desirable locations for business development, for example, suburban locations.

The results also show interesting regional variations in the transit–new business relationship: new businesses of all types in Philadelphia and Boston are most consistently associated with transit (of all modes), at 71% and 42% positive/significant coefficients respectively. While the results of this analysis do not lend significant insight into its underlying causes, it is possible that the more extensive, connected, mature – and thus useful and visible – nature of the transit networks in these regions increases their attractiveness for new business development. Historic, walkable land-use patterns that make everyday access to transit easier also may play a role. These findings also begin to shed some light on the question of the ‘novelty factor’ for transit development – it appears that the association between transit proximity and new business development can remain significant long after construction in regions with long-established, mature transit systems.

Given these results, the final question of relevance is how the pooled and regional spatial Durbin models presented here compare with the type of non-spatial Poisson regressions commonly used in the literature to this point (Chatman et al., 2016; Credit, 2017). To do this, pooled and regional Poisson regressions that approximate the spatial models specified in equation (6) for each of the four dependent variables of interest are run; for full results, see Appendix C in the supplemental data online. These models (by design) use the raw count of new businesses per block with acres as the offset and include a lagged dependent variable to control for lag-type spatial autocorrelation.

The signs and significances (at $p \leq 0.05$) for each of the transit variables in both sets of models are then compared in Table 6. If, for instance, the 0.25–0.5-mile light-rail transit variable for knowledge businesses was positive and significant in both the pooled spatial Durbin model and the Poisson model, it is marked as ‘+ Both’; if both models indicated negative significant

Table 5. Pooled and region-specific spatial Durbin model significant total effects for mode and business type.

Business type	Rail mode	Distance	Pooled	San Jose	Austin	Cleveland	Philadelphia	Boston	Total regional			Overall average by model	
									Positive	%			
Knowledge	Light rail	0.25	+			-			0	13%	41%	Light rail	22%
		0.50					+		1				
	Heavy rail	0.25						+	1	33%		Heavy rail	25%
		0.50	+					+	1				
Commuter rail	0.25	+	+				+	+	3	75%		Commuter rail	56%
	0.50	+		+			+	+	3				
High-technology	Light rail	0.25							0	0%	27%		
		0.50							0				
	Heavy rail	0.25	+					+	1	33%			
		0.50	+					+	1				
Commuter rail	0.25	+					+	+	2	50%			
	0.50	+					+	+	2				
Retail, food and services	Light rail	0.25	+	-				+	+	2	38%	45%	
		0.50	+					+	1				
	Heavy rail	0.25	+					+	1	33%			
		0.50	+					+	1				
Commuter rail	0.25	+	+				+	+	3	63%			
	0.50	+					+	+	2				

Producer services	Light rail	0.25	+			-	+	+	2	38%	27%
		0.50				+			1		
	Heavy rail	0.25							0	0%	
		0.50							0		
	Commuter rail	0.25	+				+	+	2	38%	
		0.50	+					+	1		
	Total		Positive		2	1	1	17	10		
	regional		Percent		13%	13%	6%	71%	42%		

Table 6. Differences between spatial Durbin and Poisson model results, i.e., Type I and II errors.

		Corroboration			Type I error		Type II error		Unknown
		+ Both	– Both	Not significant	+ Poisson	– Poisson	+ Durbin	– Durbin	Mixed
Rail mode	Light rail	11	0	1	25	0	0	0	3
	Heavy rail	11	0	2	19	0	0	0	0
	Commuter rail	26	0	0	14	0	0	0	0
Business type	Knowledge	13	0	0	14	0	0	0	1
	High-tech	10	0	2	16	0	0	0	0
	Retail	16	0	1	10	0	0	0	1
	Producer services	9	0	0	18	0	0	0	1
Region	Pooled	17	0	0	7	0	0	0	0
	San Jose	2	0	1	12	0	0	0	1
	Austin	1	0	0	7	0	0	0	0
	Cleveland	1	0	2	11	0	0	0	2
	Philadelphia	17	0	0	7	0	0	0	0
	Boston	10	0	0	14	0	0	0	0
Total		48	0	3	58	0	0	0	3
%			46%		52%		3%		3%

coefficients, that variable was marked as ‘– Both’, and ‘Not sig.’ for variables insignificant in both sets of models. More interesting than these corroborations, however, are the instances where the Poisson models show a positive/significant result while the variable is not significant in the spatial models (+ Poisson’); given the increased information available in the spatial models, we can confidently think of these instances as Type I errors; Type II errors, then, are instances where the spatial models show a positive/significant result that is not captured in the Poisson models (+ Durbin’, which occurs much less frequently). There are also a small number of instances where both types of models show significant results with opposite signs, which are categorized as ‘Mixed’ results.

While there is some corroboration between the spatial and non-spatial approaches (46%), the results show that 52% of the positive/significant coefficients that appear in non-spatial Poisson models are Type I errors, while 3% are Type II errors. Table 6 shows this percentage broken down by mode, business type and region – each of those (large) rows conveys the same total counts but shows how the Type I and Type II errors are distributed across those different categories of interest. Perhaps most interestingly, in San Jose, Austin, Cleveland and Boston, the number of Type I errors is larger than the number of corroborations, indicating that non-spatial models may significantly overestimate the relationship between transit and new business creation in general, and specifically in regions with less mature transit systems.

DISCUSSION AND CONCLUSIONS

Based on these results, three primary conclusions can be drawn about the relationship between new businesses and transit in a variety of regions. First, the results show that proximity to rail transit stations does, in fact, have a positive overall relationship with new business starts, even while controlling for several forms of spatial dependence, total existing and new business activity in the block, and other socio-demographic factors. New retail, services, and food business and knowledge businesses are most consistently associated with rail transit variables. At the same time, commuter rail shows the most consistent association with adjacent new business creation of all types, and regions with extensive transit networks and a dense, historic urban fabric – Philadelphia and Boston – show the most consistent association between new business creation and transit proximity.

Second, empirical testing of the performance of different possible ‘exposure’ variables for new business creation – including area, total population, existing business activity and all new business activity – shows that area provides the most consistent, stable foundation for calculating expected rates of new business activity. It also provides the most easy-to-interpret measure; talking about ‘expected new business density’ is much clearer and easy to visualize than ‘expected new business creation out of all existing businesses’, for example. As for smoothing functions, spatial smoothers (including spatial rate and SEB) create the most effective estimates of expected new business activity, balancing coverage (expanding the number of observations) with model performance. Testing a range of spatial econometric models also indicates that the spatial Durbin model provides the lowest Akaike information criterion (AIC) specification.

Third, this paper shows that spatial econometric models are necessary for estimating the relationship between aggregated new business starts and transit proximity – comparisons made between spatial model results and similar non-spatial Poisson models indicate that non-spatial models overestimate the significance of rail transit proximity 52% of the time. These Type I errors are particularly common in regions with less established transit networks. This indicates that previous work on the new business – transit connection may significantly overestimate the effect of transit proximity, perhaps with negative policy implications.

Of course, this paper has several limitations as well. It does not include a direct test of other approaches to modelling spatial count data, including spatial GLMM and spatial filtering – a

useful extension of this work could directly compare the performance of those approaches with the current one (in terms of computation time, control for spatial dependence, bias in parameter estimates, etc.) in order to see which is most useful for regional science practitioners and researchers. In addition, this paper does not capture causal relationships or evaluate trends over time; future work could expand the approach used here with panel data in order to evaluate better whether the observed relationships change over time.

Despite these limitations, this analysis provides a useful addition to the literature on the economic impacts of transit. While previous work has shown a strong connection between knowledge business creation and transit proximity for relatively new light-rail systems (Chatman et al., 2016; Credit, 2017), this paper shows that transit proximity also has a significant positive relationship with a range of new business types. These findings could have important implications for urban planners and policy-makers evaluating the economic costs and benefits of creating or extending new rail transit systems. New transit systems can certainly play a role in catalyzing new business development but appears more likely to do so if the systems are extensive, connect employment centres and are supported by walkable urban environments.

Equally interesting are the implications of this paper for future research. By empirically testing different exposure variables for new business creation, this analysis shows that area provides the most stable background from which to calculate new business creation rates. In a theoretical sense, this also suggests that new business creation is a process influenced more by density than demand (population) or co-location with (all) other types of businesses. This paper also provides a novel method for modelling spatial count data within traditional spatial econometric frameworks and demonstrates the importance of using spatial models in transit-new business research; without such models, the relationship between transit and new business creation is likely to be vastly overestimated.

DISCLOSURE STATEMENT

No potential conflict of interest was reported by the author.

FUNDING

This work was supported by the Ewing Marion Kauffman Foundation [grant number 20130782]. It was also supported by the Graduate School at Michigan State University under a Dissertation Completion Fellowship.

SUPPLEMENTAL DATA

Supplemental data for this article can be accessed at <https://doi.org/10.1080/17421772.2019.1523548>

NOTES

¹ Hecker provides an often-cited definition of high-technology employment based on NAICS industries with high concentrations of high-tech employment. The 14 industries identified as 'Level 1' are (by NAICS code): 3254, 3341, 3342, 3344, 3345, 3364, 5112, 5161, 5179, 5181, 5182, 5413, 5415 and 5417 (Hecker, 2005). For the purposes of this paper, new businesses in these sectors are classified as high-tech start-ups.

² Transformed, as described in the Methods section below, for use in the final model.

³ This is the terminology used in GeoDa v.1.8.16.4 1, which was used to calculate all the smoothed variables used in the paper.

⁴ The negative binomial model is a special case of the Poisson model (with an additional parameter) that is used to model overdispersion in a Poisson distribution, that is, when the variance of the distribution is greater than its mean (Hilbe, 2014).

⁵ Note that this paper does not attempt to argue that the method used here is statistically preferable to either the spatial GLMM or the spatial filtering approaches for modeling spatial count data – a direct comparison of these methods is beyond the scope of this paper. Rather, this study simply presents an alternative method for modeling business data within traditional spatial econometric frameworks.

⁶ A first-order queen spatial weights matrix was used to calculate the spatial rate and SEB smoothing rates.

⁷ Many thanks to Dr LeSage for providing the MATLAB code to perform this analysis on his website (see <http://www.spatial-econometrics.com>) and via email. These posterior model probabilities were calculated using the *lmarginal_cross_section* function using the MATLAB code outlined in LeSage (2015).

⁸ All final models were estimated using the ‘lagsarlm’ function in the ‘spdep’ R package using the Monte Carlo approximate log-determinant method of weights matrix decomposition.

⁹ Significance measured at the $p \leq .05$ level.

ORCID

Kevin Credit  <http://orcid.org/0000-0002-3320-4670>

REFERENCES

- Agostini, C. A., & Palmucci, G. (2008). The anticipated capitalisation effect of a New metro line on housing prices. *Fiscal Studies*, 29, 233–256. doi:10.1111/j.1475-5890.2008.00074.x
- Aharonson, B. S., Baum, J. A. C., & Feldman, M. P. (2007). Desperately seeking spillovers? Increasing returns, industrial organization and the location of new entrants in geographic and technological space. *Industrial and Corporate Change*, 16, 89–130. doi:10.1093/icc/dtl034
- Aharonson, B. S., Stettner, U., Amburgey, T. L., Ellis, S., & Drori, I. (2013). *Understanding the relationship between networks and technology, creativity and innovation*. Bingley, UK: Emerald Group Publishing Limited.
- American Public Transit Association (APTA). (1994). Glossary of Transit Terminology. American Public Transit Association. Retrieved from http://www.apta.com/resources/reportsandpublications/Documents/Transit_Glossary_1994.pdf
- Anselin, L. (1988). *Spatial econometrics: Methods and models*. Dordrecht: Kluwer.
- Anselin, L., Kim, Y. W., & Syabri, I. (2004). Web-based analytical tools for the exploration of spatial data. *Journal of Geographical Systems*, 6, 197–218. doi:10.1007/s10109-004-0132-5
- Anselin, L., & Rey, S. (2014). *Modern spatial econometrics in practice*. Chicago, IL: GeoDa Press.
- Batty, M., Besussi, E., & Chin, N. (2003). Traffic, Urban Growth, and Suburban Sprawl. *University College London Working Paper Series: Paper 70*. Retrieved from <https://www.bartlett.ucl.ac.uk/casa/pdf/paper70.pdf>
- Bollinger, C. R., & Ihlanfeldt, K. R. (1997). The impact of rapid rail transit on economic development: The case of Atlanta’s MARTA. *Journal of Urban Economics*, 42, 179–204. doi:10.1006/juec.1996.2020
- Cervero, R. (1984). Journal report: Light rail transit and urban development. *Journal of the American Planning Association*, 50(2), 133–147. doi:10.1080/01944368408977170
- Cervero, R. (1994). Rail transit and joint development: Land market impacts in Washington, DC, and Atlanta. *Journal of the American Planning Association*, 60(1), 83–94. doi:10.1080/01944369408975554

- Cervero, R. (2004). Effects of light and commuter rail transit on land prices: Experiences in San Diego County. *Journal of the Transportation Research Forum*, 43(1), 121–138.
- Chapple, K., Markusen, A., Schrock, G., Yamamoto, D., & Yu, P. (2004). Gauging metropolitan ‘high-tech’ and ‘I-tech’ activity. *Economic Development Quarterly*, 18(1), 10–29. doi:10.1177/0891242403257948
- Chatman, D. G. (2013). Does TOD need the T? *Journal of the American Planning Association*, 79(1), 17–31. doi:10.1080/01944363.2013.791008
- Chatman, D. G., & Noland, R. B. (2011). Do public transport improvements increase agglomeration economies? A review of literature and an agenda for research. *Transport Reviews*, 31(6), 725–742. doi:10.1080/01441647.2011.587908
- Chatman, D. G., Noland, R. B., & Klein, N. J. (2016). Firm births, access to transit, and agglomeration in Portland, Oregon, and Dallas, Texas. *Transportation Research Record: Journal of the Transportation Research Board*, 2598, 1–10. doi:10.3141/2598-01
- Christensen OF, Roberts GO, Sköld M (2006) Robust Markov chain Monte Carlo methods for spatial generalized linear mixed models. *Journal of Computational and Graphical Statistics* 15(1): 1–17. doi:10.1198/106186006X100470
- Clayton, D., & Kaldor, J. (1987). Empirical Bayes estimates of age-standardized relative risks for Use in disease mapping. *Biometrics*, 43, 671–681. doi:10.2307/2532003
- Credit, K. (2017). Transit-oriented economic development: The impact of light rail on new business starts in the phoenix, AZ region. *Urban Studies*. doi:10.1177/0042098017724119
- Damm, D., Lerman, S. R., Lerner-Lam, E., & Young, J. (1980). Response of urban real estate values in anticipation of the Washington metro. *Journal of Transport Economics and Policy*, 14(3), 315–336.
- da Silva, D. F. C., Elhorst, J. P., & Neto, R. (2017). Urban and rural population growth in a spatial panel of municipalities. *Regional Studies*, 51(6), 894–908. doi:10.1080/00343404.2016.1144922
- De Oliveira, V. (2013). Hierarchical Poisson models for spatial count data. *Journal of Multivariate Analysis*, 122, 393–408. doi:10.1016/j.jmva.2013.08.015
- DeVol, R., & Wong, P. (1999). *America’s high-tech economy: Growth, development and risks for metropolitan areas*. Santa Monica, CA: Milken Institute.
- Diggle, P. J., & Ribeiro, P. J. (2007). *Model-based geostatistics*. New York, NY: Springer.
- Diggle, P. J., Tawn, J. A., & Moyeed, R. A. (1998). Model-based geostatistics. *Applied Statistics*, 47(3), 299–350.
- Elhorst, J. P. (2014). *Spatial econometrics: From cross-sectional data to spatial panels*. New York, NY: Springer.
- Elhorst, J. P. (2017). Spatial panel data analysis. In S. Shekhar, H. Xiong, & X. Zhou (Eds.), *Encyclopedia of GIS*, (2nd ed.), 2050–2058. Cham, Switzerland: Springer International Publishing.
- Elhorst, J. P., Abreu, M., Amaral, P., Bhattacharjee, A., Corrado, L., Fingleton, B., ... Yu, J. (2016). Raising the bar (1). *Spatial Economic Analysis*, 11(1), 1–6. doi:10.1080/17421772.2015.1126966
- Ewing, R., & Cervero, R. (2010). Travel and the built environment: A meta-analysis. *Journal of the American Planning Association*, 76(3), 265–294. doi:10.1080/01944361003766766
- Faraway, J. J. (2006). *Extending the linear model with R: Generalized linear, mixed effects and nonparametric regression models*. New York: Chapman & Hall.
- Freemark, Y. (2014). Have U.S. Light Rail Systems Been Worth the Investment? *Citylab*. Retrieved from <http://www.citylab.com/commute/2014/04/have-us-light-rail-systems-been-worth-investment/8838/>
- Gerkman, L. M., & Ahlgren, N. (2014). Practical proposals for specifying k-nearest neighbours weights matrices. *Spatial Economic Analysis*, 9(3), 260–283. doi:10.1080/17421772.2014.930167
- Gilbert, B. A., McDougall, P. P., & Audretsch, D. (2008). Clusters, knowledge spillovers and new venture performance: An empirical examination. *Journal of Business Venturing*, 23(4), 405–422. doi:10.1016/j.jbusvent.2007.04.003
- Giuliano, G. (2004). Land Use impacts of transportation investments: Highway and transit. In S. Hanson, & G. Giuliano (Eds.), *The geography of urban transportation* (pp. 237–273). New York: Guilford Press.
- Golub, A., Guhathakurta, S., & Sollaapuram, B. (2012). Spatial and temporal capitalization effects of light rail in phoenix: From conception, planning, and construction to operation. *Journal of Planning Education and Research*, 32(4), 415–429. doi:10.1177/0739456X12455523

- Hecker, D. E. (2005). High-technology employment: A NAICS-based update. *Monthly Labor Review*, 128(7), 57–72.
- Hilbe, J. (2014). *Modeling count data*. Cambridge, MA: Cambridge University Press.
- Jing, L., & De Oliveira, V. (2015). Geocount: An R package for the analysis of geostatistical count data. *Journal of Statistical Software*, 63(11), 1–33. doi:10.18637/jss.v063.i11
- Kafadar, K. (1996). Smoothing geographical data, particularly rates of disease. *Statistics in Medicine*, 15, 2539–2560. doi:10.1002/(SICI)1097-0258(19961215)15:23<2539::AID-SIM379>3.0.CO;2-B
- Knaap, G., Ding, J. C., & Hopkins, L. D. (2001). Do plans matter? The effects of light rail plans on land values in station areas. *Journal of Planning Education and Research*, 21(1), 32–39. doi:10.1177/0739456X0102100103
- Knight, R. L., & Trygg, L. L. (1977). Evidence of land Use impacts of rapid transit systems. *Transportation*, 6(3), 231–247. doi:10.1007/BF00177453
- Landis, J., Guhathakurta, S., & Zhang, M. (1994). Capitalization of transit investments into single-family home prices: A comparative analysis of five California rail transit systems. *The University of California Transportation Center*, 246, 1–38. Retrieved from <http://www.uctc.net/papers/246.pdf>
- Lawson, A. B., Banerjee, S., Haining, R. P., & Ugarte, M. D. (2016). *Handbook of spatial epidemiology*. New York, NY: CRC Press.
- LeSage, J. P. (2014). Spatial econometric panel data model specification: A Bayesian approach. *Spatial Statistics*, 9, 122–145. doi:10.1016/j.spasta.2014.02.002
- LeSage, J. P. (2015). Software for Bayesian cross section and panel spatial model comparison. *Journal of Geographical Systems*, 17, 297–310. doi:10.1007/s10109-015-0217-3
- LeSage, J. P., & Pace, R. K. (2009). *Introduction to spatial econometrics*. New York, NY: CRC Press.
- Liu, H., Davidson, R. A., & Apanasovich, T. V. (2007). Statistical forecasting of electric power restoration times in hurricanes and Ice storms. *IEEE Transactions on Power Systems*, 22(4), 2270–2279. doi:10.1109/TPWRS.2007.907587
- Mack, E., & Credit, K. (2016). The intra-metropolitan geography of entrepreneurship: A spatial, temporal, and industrial analysis (1989–2010). In E. Mack, & H. Qian (Eds.), *Geographies of entrepreneurship* (pp. 101–121). New York: Taylor and Francis.
- Malecki, E. J. (1984). High technology and local economic development. *Journal of the American Planning Association*, 50(3), 262–269. doi:10.1080/01944368408976593
- Mejia-Dorantes, L., Paez, A., & Vassallo, J. M. (2012). Transportation infrastructure impacts on firm location: The effect of a new metro line in the suburbs of Madrid. *Journal of Transport Geography*, 22, 236–250. doi:10.1016/j.jtrangeo.2011.09.006
- Mohammad, S., Graham, D., Melo, P., & Anderson, R. (2013). A meta-analysis of the impact of rail projects on land and property values. *Transportation Research Part A*, 50, 158–170.
- O’Hullachain, B., & Reid, N. (1991). The location and growth of business and professional services in American metropolitan areas, 1976–1986. *Annals of the Association of American Geographers*, 81(2), 254–270. doi:10.1111/j.1467-8306.1991.tb01689.x
- Quinn, B. (2006). Transit-oriented development: Lessons from California. *Built Environment*, 32(3), 311–322. doi:10.2148/benv.32.3.311
- Renski, H. (2008). New firm entry, survival, and growth in the United States: A comparison of urban, suburban, and rural area. *Journal of the American Planning Association*, 75(1), 60–77. doi:10.1080/01944360802558424
- Richardson, H., & Bae, C. (2016). *Urban sprawl in Western Europe and the United States*. New York, NY: Ashgate Publishing.
- Richman, J. (2015). Here’s Why Tech Startups are Flocking to Austin. *Tech.co*. Retrieved from <https://tech.co/make-way-silicon-valley-heres-tech-startups-flocking-austin-2015-04>
- Rios, V., Pascual, P., & Cabases, F. (2017). What drives local government spending in Spain? A dynamic spatial panel approach. *Spatial Economic Analysis*, 12(2–3), 230–250. doi:10.1080/17421772.2017.1282166
- Saxenian, A. (1994). *Regional advantage: Culture and competition in silicon valley and route 128*. Cambridge, MA: Harvard University Press.

- Schuetz, J. (2014). Do rail transit stations encourage neighborhood retail activity? *Urban Studies*. doi:10.1177/0042098014549128
- Song, Y., Lee, K., Anderson, W. P., & Lakshmanan, T. R. (2012). Industrial agglomeration and transport accessibility in metropolitan Seoul. *Journal of Geographical Systems*, 14(3), 299–318. doi:10.1007/s10109-011-0150-z
- Squires, G. (2002). *Urban sprawl: Causes, consequences, and policy responses*. Washington, DC: The Urban Institute Press.
- Tiefelsdorf, M., & Griffith, D. A. (2007). Semiparametric filtering of spatial autocorrelation: The eigenvector approach. *Environment and Planning A*, 39(5), 1193–1221. doi:10.1068/a37378
- Vessali, K. V. (1996). Land Use impacts of rapid transit: A review of the empirical literature. *Berkeley Planning Journal*, 11(1), 72–105.
- Walls & Associates. (2012). NETS Database vs. BLS Establishment & Employment Estimates. *Walls & Associates: Insights into Business Dynamics*. Obtained from NETS datafile.
- Wang, Y., Kockelman, K., & Wang, X. (2013). Understanding spatial filtering for analysis of land use-transport data. *Journal of Transport Geography*, 31, 123–131. doi:10.1016/j.jtrangeo.2013.06.001
- Weinberger, R. R. (2001). Light rail proximity: Benefit or detriment in the case of Santa Clara county, California? *Transportation Research Record: Journal of the Transportation Research Board*, 1747, 104–111. doi:10.3141/1747-13
- Weinstein, B., & Clower, T. L. (2003). DART light rail's effect on taxable property valuations and transit-oriented development. *University of North Texas*. Retrieved from http://www.valleymetro.org/images/uploads/general_publications/2003_DART_Study.pdf
- Weisbrod, G., Duncan, C., & Moses, S. (2014). Evolving connection of transit, agglomeration, and growth of high-technology business clusters. *Transportation Research Record: Journal of the Transportation Research Board*, 2452, 11–18. doi:10.3141/2452-02