geographical analysis

Geographical Analysis (2021) **0**, 1–26

Spatial Models or Random Forest? Evaluating the Use of Spatially Explicit Machine Learning Methods to Predict Employment Density around New Transit Stations in Los Angeles

Kevin Credit

Center for Spatial Data Science, University of Chicago, Chicago, IL USA

The increasing use of "new" machine learning techniques, such as random forest, provides an impetus to researchers to better understand the role of space in these models. Thus, this article develops an approach for constructing spatially explicit random forest models by including spatially lagged variables to mirror various spatial econometric specifications in order to test their comparative performance against traditional spatial and nonspatial regression models for predicting block-level employment density around new transit stations in Los Angeles. This article employs a "post hoc" testing approach to isolate the impact of a particular variable (transit proximity)—and supplemental diagnostics (such as partial dependence plots and permutation importances)—to help inform explanatory relationships. The results indicate that random forest models slightly outperform spatial econometric models, and the inclusion of spatial lag parameters modestly improves random forest model accuracy—the best-fit spatial random forest model demonstrates 84.61% accuracy in predicting post-construction employment density around newly built transit stations, compared to 81.88% for the best-fit spatial econometric model and 84.37% for the nonspatial random forest model. However, given these somewhat small differences, it is not possible to conclude that the random forest approach is clearly superior to traditional spatial econometric models from these results alone.

Introduction

For urban and economic geographers, the last 20 years can be characterized by a rapid increase in the number of large spatially referenced data sets, both publicly and privately provided. These include fine-grained data from the U.S. Census such as the Longitudinal Employer-Household Dynamics Origin-Destination Employment Statistics (LODES), individual business establishment data such as the National Establishment Time Series (NETS), and large open data sets provided by cities on their open data portals, such as the Transportation Network Providers (TNP) data set on individual rideshare trips from the City of Chicago.

Correspondence: Kevin Credit, Center for Spatial Data Science, University of Chicago, 5735 S. Ellis Ave., Room 230, Chicago, IL 60637, USA
e-mail: kcredit@uchicago.edu

Submitted: December 28, 2019. Revised version accepted: December 15, 2020.

doi: 10.1111/gean.12273

At the same time, a variety of new methods have emerged from the burgeoning field of "data science" that are particularly suited to evaluating these large data sets, including random forests (RF) and artificial neural networks (ANN) (Openshaw and Openshaw 1997; Breiman 2001a). Data science has gained prominence as an approach within statistics that focuses on using empirical relationships to build theory rather than relying on predetermined assumptions (Breiman 2001b), fostered in large part by this new combination of large data sets and computationally powerful methods to analyze them. However, several issues have emerged as the new data science has begun to interact more directly with spatially explicit data and research questions: (1) these new methods, often developed in the private sector and generally focused on predictive applications, are not designed to expose the explanatory relationships between variables that geographers are typically interested in, and thus are still seen as "black boxes" by many researchers; (2) the role of space in these models is still poorly understood, and the inclusion of spatially explicit relationships in these models has not yet been rigorously assessed; and (3) these methods are often employed without any prior theoretical knowledge of the spatial or substantive relationships that might be driving or informing the results (Rey 2019; Singleton and Arribas-Bel 2019).

This article seeks to begin to solve some of these issues by answering the following questions: first, how do predictive machine learning (in this case, random forest) models compare to traditional spatial econometric (SE) techniques in terms of prediction accuracy? What factors are associated with higher performance for random forest versus traditional spatial econometric models? And, from a spatial perspective, do spatially explicit random forest models outperform standard (nonspatial) random forest models?

These questions are assessed through an application: evaluating the adjacent employment impacts of the construction of a new transit line in Los Angeles County. The LA Metro has, as of 2019, the third-highest ridership of any light rail system in the United States (APTA), with six lines constructed between 1990 and 2016. The first phase of the Expo Line opened in 2012, while Phase II to Santa Monica opened in 2016, and thus provides a useful test case for the predictive capabilities of these models. The application is also substantively relevant, given the historic and continuing research attention devoted to evaluating the economic development impacts of transit proximity (Knight and Trygg 1977; Damm et al. 1980; Cervero 1984; Green and James 1993; Cervero 1994; Landis, Guhathakurta, and Zhang 1994; Cervero and Landis 1997; Bollinger and Ihlandfeldt 1997; Knaap, Ding, and Hopkins 2001; Weinberger 2001; Weinstein and Clower 2003; Cervero 2004; Hess and Almeida 2007; Agostnini and Palmucci 2008; Golub, Guhathakurta, and Sollapuram 2012; Mohammed et al. 2013; Seo, Golub, and Kuby 2014; Chatman, Noland, and Klein 2016; Credit 2018) and recent findings that the use of spatial methods are vital to understanding the true effect of transit proximity on economic development (Credit 2019). While newer studies primarily focus on transit impacts to property values and new business creation (rather than employment[1]), the use of LODES employment counts at the block level as the dependent variable of interest here provides an example grounded in one of the most useful large, spatially referenced open source data sets that can easily be exported to other research contexts.

Training sets for each model are built using 2010 data; the predictions of these 2010 models are then tested on the actual employment density of blocks within 800 m of new Expo Line stations in the time after the line opened. Differences in error between models are then tested against various explanatory variables in order to better understand the factors which contribute to an increase in explanatory power for one model versus another. This article also lays out a

method for constructing spatially explicit RF models by including spatially lagged variables in a way that mirrors various SE specifications (such as the spatial lag and spatial Durbin models).

The results of this analysis indicate that random forest models modestly outperform traditional spatial econometric models in each case, and that the inclusion of spatial parameters incrementally increases the predictive accuracy of the random forest models. However, the increase in predictive accuracy for the best-fit spatial random forest model compared to the nonspatial random forest model is somewhat minimal (84.61% versus 84.37%, respectively), and while the improvement over the best-fit spatial econometric specification (81.88%) is more substantial, it is not possible to conclude that the random forest approach is *clearly* superior to traditional spatial econometric models from these results alone.

Still, despite the relatively small size of the differences demonstrated here, these findings are important because they do show better baseline performance for the spatial random forest model (which could likely be improved even further in less traditionally parsimonious specifications[2]), and the spatial lag parameters also demonstrate high (permutation) importance to the random forest Durbin model (up to 19%), suggesting that the inclusion of spatially explicit variable constructions may be a fruitful approach to explore as work continues on the development of more spatially explicit machine learning methods.

While these results are inconclusive on the question of model superiority, they are still useful for (1) scholars interested in better incorporating space explicitly into newer machine learning models (e.g., Singleton and Arribas-Bel 2019), and also for (2) exploring how study design—in this case, the "post hoc" testing approach that isolates the impact of a particular variable (transit proximity)—and supplemental diagnostics (such as partial dependence plots and permutation importances) can help inform the explanatory relationships that geographers and other social scientists are traditionally interested in.

While this article is primarily focused on a predictive application, it also provides a first step for thinking about how random forest—and possibly other, newer machine learning techniques—might be employed to assess explanatory relationships and research questions in a spatial context. Given the ongoing development of approaches that employ the random forest framework to make explanatory inferences, such as "causal" trees (Athey and Imbens 2016), it appears that these newer machine learning methods are poised to drive the cutting edge of quantitative data analysis in the near future. While it is not yet clear that these "new" methods are unequivocally superior to traditional spatial econometric techniques, this article provides a useful contribution to the burgeoning literature on spatially explicit machine learning methods by charting out the initial territory for understanding the spatial implications—and uses—of these methods.

Literature

Before exploring the existing literature on machine learning in urban geography, a few definitions are instructive. Conventionally, "machine learning" methods are classified as a branch of the larger field of artificial intelligence (AI) that "learn from data" to "perform predictions on unknown data" (Prateek 2017, p. 14). These are further categorized into "supervised" and "unsupervised" methods—supervised algorithms are trained using both independent (X) and dependent (y) variables and used to predict new instances based on the examined relationship between X and y (known as "out of sample" prediction). Supervised methods include many familiar statistical techniques such as linear and logistic regression, as well as newer techniques such as support vector machines (SVM), decision trees, random forest, and neural networks (Géron

2017). Unsupervised algorithms, on the other hand, produce a predicted *y* (often a classification) based only on the relationships between the provided *X* variables, as in clustering techniques such as k-means or dimensionality reduction techniques such as principle component analysis (PCA) (Géron 2017).

These distinctions make sense when viewed from the perspective of *predictive* applications but can be confusing to those who have used many of these statistical techniques for years in *explanatory* applications. Of course, for urban and economic geographers, *explanatory* or *inferential* applications—in which models explicitly quantify and produce the statistical relationship between *X* and *y*—are often of primary importance. In addition, while many (generally conventional) machine learning algorithms produce measures of these relationships (e.g., linear or logistic regression), other (generally newer) algorithms do not (e.g., decision trees, Random Forests or neural networks), and thus are often criticized or ignored as "black box" methods.[3] For this reason, the more important distinction for geographers is between the primarily *predictive* machine learning methods (such as RF) that generally do not produce coefficients or measures of the relationship between *X* and *y* and the more conventional *explanatory* machine learning methods that many urban and economic geographers are familiar with (such as linear regression). Of course, explanatory methods can also be used to make predictions, and (as this article's application shows) some explanatory features can be ascertained when using predictive models (such as RF).

Although Openshaw (Openshaw and Openshaw 1997) advocated for the adoption of predictive AI methods in geography, their use in studying urban geographic problems has accelerated only in recent years (Grekousis 2019) with the development of more accessible, easy-to-use implementations such as the *sci-kit learn* package in Python. Even with this recent growth, the majority of urban-geographic applications of predictive models come from remote sensing—for instance, a recent meta-analysis of the use of artificial neural networks in urban geography found that only 15% of the eligible studies examined "socioeconomic" topics (as opposed land cover/land use or urban environmental issues), and that the vast majority of articles analyzed used satellite data and were published in remote sensing journals (Grekousis 2019).

While image classification is one of the most prominent predictive applications in urban geography, there are of course other important predictive questions that can be answered in the era of "big" data: small area estimation and interpolation for socioeconomic data (Singleton and Arribas-Bel 2019), spatial patterns in large, open, georeferenced municipal data sets such as crimes, "311" calls, and parking violations (Gao et al. 2019), spatiotemporal patterns in disease outbreaks using georeferenced sentiment data from social media (e.g., Allen et al. 2016), the spatial distribution of pollution (Walsh et al. 2017), the prediction of housing prices and rents (Mu, Wu, and Zhang 2014; Fan, Cui, and Zhong 2018; Phan 2018; Truong et al. 2020), and gentrification (Alejandro and Palafox 2019; Knorr 2019), among others. In an urban planning context, predicting the future distribution of population and land use with greater precision is an area of significant opportunity for predictive model applications (Feng et al. 2018). Indeed, while this article's application is concerned primarily with predicting employment density around transit, the methods delineated here could be used to predict regional (workplace-level) employment and residential population growth more generally.

In addition to the application of existing predictive machine learning methods to urban geographic questions, scholars have identified the need to create *spatially explicit* predictive models and methods (Janowicz et al. 2019; Singleton and Arribas-Bel 2019). As Singleton & Arribas-Bel (2019, p. 9) concisely point out, "one of the most fruitful methodological areas

where Geographic Data Science could comprehensively rework some of those core techniques of Data Science" is by explicitly including space to improve the performance of predictive machine learning models. Indeed, the small number of existing studies that create spatially explicit prediction frameworks tend to show that these methods perform better than nonspatial methods when applied to spatial data sets (Hengl et al. 2018; Georganos et al. 2019; Janowicz et al. 2019; Yan et al. 2019). To better understand the role of space in predictive machine learning models, this article systematically tests the use of various spatial lags in a RF model, compares RF models to more conventional spatial econometric models, and evaluates the factors that lead to better predictive performance for the RF or spatial econometric models.

"Post hoc" testing approach

The general approach used in this article is to test the accuracy of various spatial econometric and random forest model specifications in predicting post-construction employment for a new transit line at the block level by first building a baseline 2010 model to train the relationship between expected employment density (explained in more detail below) and a small number of parsimonious covariates, including a dummy for location within 800 m of a transit station existing in 2010 (i.e., not the Expo Line, since it was not yet built).

These relationships were then applied to two testing scenarios, with two tested y variables: (1) the pre-construction observed 2010 expected employment density for a random 20% subset of the data (standard practice in machine learning model validation), and (2) observed post-construction expected employment density for blocks within 800 m of the new Expo Line (with the transit dummy updated in these blocks to equal 1 to reflect their new status). Scenario 1 is a completely cross-sectional prediction, with no consideration for pre-/post-construction dynamics, using 2010 variables to find a traditional measure of prediction accuracy based on the overall mean absolute percentage error (MAPE) between predicted values of expected 2010 employment density and observed values of expected 2010 employment density. In Scenario 1, out-of-sample validation is accomplished by training the models on a random 80% subset of the data and predicting values to the remaining (out-of-sample) 20% of the data.[4]

In Scenario 2, the design of the out-of-sample prediction case is meant to test the accuracy of the predictions generated by the pre-construction relationships: the models are again trained using the variables in 2010, with one exception: blocks within 800 m of the Expo line stations are now given a 1, in order to answer the question that the article is interested in: how well does a model trained on pre-construction data predict *actual* post-construction outcomes? The idea behind this "post hoc" design is taken from thinking about a real-world prediction case: imagine there is a planning debate in LA County in 2010 about the economic effect of opening a new transit line. Planners test the existing relationship between employment and transit proximity and find a relationship (e.g., a regression coefficient), which in effect provides a prediction of what would happen to expected employment density in blocks in which a new transit line was built (i.e., with the transit dummy updated to equal 1 in those blocks).

But how well will the planners do in actually predicting the effect of transit on employment density? And, more importantly, which type of model will do better? It is impossible to know at the time, but with the benefit of hindsight it is possible to evaluate how well that (hypothetical) 2010 model *actually did* in predicting employment density around transit stations *after* the construction of the line, and to compare how different types of models did. This means that the validation for Scenario 2 is cross-sectionally *and* temporally out-of-sample—the testing *y* in

this case is the post-construction expected employment density in the blocks within 800 m of the newly built Expo line stations. This scenario uses the full (2010) data for training[5] (with the transit dummy updated to equal 1 in blocks near the new Expo line) and evaluates the predictions made by the training model against the actual observed values of post-construction employment density. Thus, the results of Scenario 2 provide insight for planners and policy-makers as to which model type will tend to *end up* providing the better prediction of real conditions in the future.

Study area and data preparation

Los Angeles County and the Expo Line

Los Angeles County was chosen as the study area for the application for two primary reasons. First, the County's Metropolitan Transportation Authority (known as Metro) is—perhaps surprisingly—among the highest-ridership light rail systems in the United States, with an average weekday ridership in the third quarter of 2019 of 140,800, placing it behind only San Francisco's Muni Railway and Boston's Massachusetts Bay Transportation Authority, two more historically established transit systems (APTA 2019). Substantively, studying the role of new transit construction on economic development in Los Angeles (the second-largest region in the United States) also makes sense, given its size and generally auto-oriented development pattern, which is more typical of most U.S. urban areas than high-density cities like Chicago or New York.

Second, the timing of the construction of the County's Expo Line fits within the constraints of the available data and the general research approach outlined above. In order to make this "true validation" approach work, the study region needed to have an existing transit system that had been operating widely for several years leading up to the baseline model training year, so that the relationships discovered by the model between transit proximity and employment were relatively well-established (i.e., somewhat insulated from short-term fluctuations such as the "novelty effect") (Mohammed et al. 2013; Credit 2018). At the same time, the study region needed to have a new line constructed after the baseline year (but within the data years available) so that the baseline effects could be applied to a newly constructed line in order to test the baseline model's true predictive capacity for a new line. As shown in Fig. 1, LA County's transit system fits these criteria perfectly. The Blue, Red, Green, Purple, and Gold Lines were all constructed between 1990 and 2003, while the Expo Line was built in two phases nearly 10 years later—Phase I, which opened in 2012, and Phase II, which opened in 2016. Thus the baseline 2010 model accounts for longstanding relationships between transit and employment in the region, and its predicted values can be aptly compared to known (and available) post-construction employment totals (2013 for Phase I and 2017 for Phase II) for station-proximate blocks.

LODES data

The data for this article come primarily from the Longitudinal Employer-Household Dynamics Origin-Destination Employment Statistics (LODES) data set form the U.S. Census Bureau. These data come from administrative records that delineate the home and work address of individual workers, such as state unemployment insurance reporting and federal worker earnings records, rather than census surveys that ask questions about commuting patterns (Graham, Kutzbach, and McKenzie 2014). LODES consists of three primary data sets at the census block level from 2002 to 2015: "Residence Area Characteristics" (RAC), which provides information on workers by their place of residence, "Workplace Area Characteristics" (WAC), which provides information

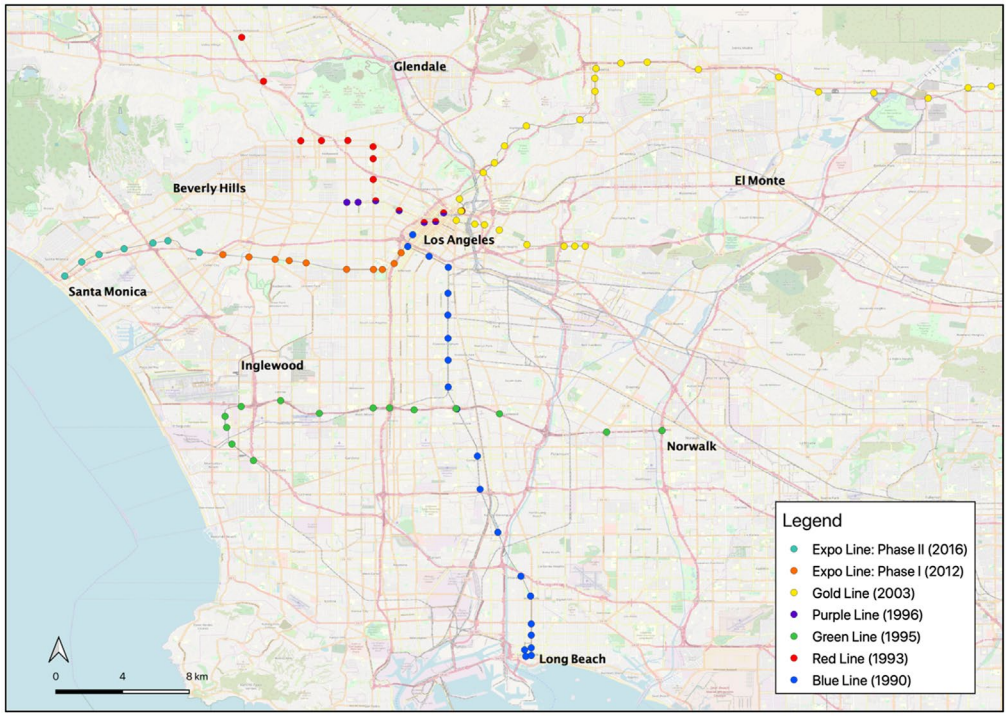


Figure 1. Map showing LA County Metro lines and years of opening.

on workers by their place of employment, and "Origin-Destination" (OD), which provides information on the individual commuting links between blocks (U.S. Census 2019).

To study the impact of new transit construction on employment density, several base variables were taken from LODES at the block level (shown, along with descriptive statistics, in Appendix Table A1): 2004 ("E_TOT04"), 2010 ("E_TOT10"), 2013, and 2017 ("POSTETOT")[6] total employment by workplace location (from WAC), 2010 employment by 2-digit NAICS code (from WAC),[7] 2010 total employment by *residential* location ("R_TOT10" from RAC), 2010 employed residents aged 29 or younger ("R_YPCT10" from RAC), and 2010 percent employed residents with earnings > $3,333/month ("R_IPCT10" from RAC). The location of transit stations was obtained from the Los Angeles County Metropolitan Transportation Agency (Metro), and blocks intersecting an 800 m buffer around these stations were identified with a transit dummy ("TRANS" for all stations and "EXPO" for Expo Line stations only).

The dependent variable—expected employment density by workplace—is one of the most widely used economic development indicators at the local level, and thus represents a baseline indicator of economic impact. The independent variables were chosen based on previous research on the economic effects of transit (Credit 2018, 2019) and both (1) control for various aspects of the economic environment that help clarify the relationship between transit and employment, and (2) (from a prediction standpoint) increase the accuracy of the employment prediction without generating spurious or over-fit results.

Several transformations of the employment variables were undertaken so that they could be used in the spatial econometric models. As previous research has shown, a spatial empirical

Bayesian (SEB) smoothing rate[8] can help correct for the instability of small counts and zero inflation when data are aggregated to a very fine spatial scale (Credit 2019). Theoretically, this data generation process in some sense "fills out" the distribution of employment density by more accurately reflecting the underlying probability of employment in a given block. In other words, because block boundaries are somewhat arbitrary—and blocks are very small—we may see a situation where, for example, one block contains 50 employees and its neighbor has 0. While this reflects the exact pattern of empirical employment based on addresses reported to unemployment insurance, etc., it does not accurately represent the spatial probability of employment, that is, the "0" block most likely does not have a true probability of employment of 0—it may certainly be lower than the 50 block, but we cannot be certain that it is 0. Thus using the SEB smoother is necessary for two reasons: (1) this study wants to preserve the finest grain of spatial variation possible in order to model the effect of transit on employment (which occurs over relatively short distances, i.e., 800 m), so the smallest spatial scale available (blocks) must be used, and (2) in order to use traditional spatial econometric models, the residuals of the model should be generally normally distributed.[9] The use of an SEB smoother (with block area as the denominator) creates an expected density measure that reduces the number of zero observations[10] and allows for log-transformation that normalizes the distribution of the dependent variable (Credit 2019).

Table 1 shows the general process for data cleaning and transformation. First, all blocks in LA County were downloaded, with those classified as "urban" retained (the blocks removed at this stage generally consist of natural areas such as mountainsides, etc.). The prepared LODES data were then joined to these blocks, with remaining missing values removed (many of these were, e.g., blocks that were made up entirely of transportation rights-of-way). At this stage, expected SEB densities for the employment variables were calculated, and log(10)-transformed; "LOGE10SEB," "LOGE04SEB," "LOGR10SEB," and "LOGEPSSEB" denote the log-transformed expected density for 2010 employment (by workplace), 2004 employment (by workplace), 2010 employment (by residence), and 1-year post-construction employment (by workplace) around the Expo Line (2013 for Phase I and 2017 for Phase II), respectively. These SEB calculations were made using a 12 nearest-neighbor spatial weights matrix; the distance matrix was chosen based on the loss of contiguity in the blocks data set from removing nonurban and missing LODES joined counts as mentioned above. Spatial lags of each of the independent and dependent variables were also calculated at this stage using a 12 nearest-neighbor spatial weights matrix. These lags were used in various RF model specifications to approximate standard spatial econometric models such as the spatial lag, Durbin, etc. Finally, all missing values at this stage were removed—since the log(10) of 0 is undefined, these observations could not be used in calculating the spatial econometric or random forest models, so in order to maintain consistency across model specifications, all missing values were simply removed. While all of the cleaning operations reduced the size of the data set by about 20% (resulting in 87,227 observations out of the raw 109,309), the resulting cleaned variables were normally distributed and able to be used in a variety of modeling applications.

Table 1. Data Cleaning Process and Number of Observations

1. All LA County blocks	109,309
2. "Urban" classification	100,319
3. Joined to LODES, retain non-zero observations	87,227
4. Calculate all SEB, log(10), and spatial lags in GeoDa using 12-nn	87,227

Methods of estimation

The general idea of this article is to test the accuracy of traditional spatial econometric models versus random forest models in the context of predicting block-level expected employment density around new transit stations. In the end, nine model specifications were tested: ordinary least squares (OLS), spatial lag (SAR), spatial error (SEM), spatial Durbin (SDM), spatial Durbin error (SDEM), random forest (RF), random forest with the spatial lag of y included (RFSAR), random forest with spatial lags of both *X* and *y* included (RFSDM), and random forest with only the spatial lag of *X* included (RFSLX). Discussion of the details of the estimation of these models follows.

Spatial econometric models

Spatial econometric models assume that the underlying data-generating process involves some form of spatial dependence—in other words, they account for the fact that "near things are more related than distant things" (Tobler 1970) by explicitly estimating parameters of spatial autocorrelation. Mathematically, spatial econometric models build on OLS specifications by inserting a measure of each observation's "neighborhood"—the spatial weights matrix (W)—directly into the model estimation process (Anselin 1988; LeSage and Pace 2009). There are a variety of specifications that place W in various positions in the standard linear regression equation, based on theoretical or data-driven concerns.

The most basic linear regression (OLS) specification is:

$$y = X\beta + u \tag{1}$$

where y is a vector of dependent variables (in this case, expected density of employment in 2010 by workplace), X is a vector of independent variables (in this case, expected density of employment in 2004 by workplace, expected density of employment in 2010 by residence, employment diversity in 2010, percent young employees in 2010 by residence, percent high income employees in 2010 by residence, and a dummy variable for location within 800 m of a transit station), β is a vector of estimated regression coefficients for these variables, and u is the error term. If spatial dependence exists in the underlying data, the OLS regression coefficients will be biased and/or the error term will be enlarged; in either case, this results in an imprecise estimation of the underlying relationships between the variables.

Spatial dependence can be explicitly modeled in a variety of ways. The spatial autoregressive (SAR) or spatial lag model inserts a parameter that captures spatial autocorrelation in the dependent variable, that is,

$$y = \rho Wy + X\beta + u \tag{2}$$

where W is a spatial weights matrix that captures the spatial neighborhood of each observation (in this case, a 12 nearest-neighbor weights matrix[11]) and ρ is the spatial autoregressive parameter on the dependent variable. However, spatial dependence could be present in the error term rather than in the dependent variable (e.g., omitted variables or spatial configuration of the data may cause the error terms of individual observations to be spatially correlated). In this case—the spatial error model (SEM)—the error term is decomposed into a spatially structured component and a random component:

$$y = X\beta + \lambda Wu + \varepsilon \tag{3}$$

where λ is the spatial autoregressive parameter in the error term and ε is the remaining unexplained error. The spatial Durbin model (SDM) is similar to the SAR model, but adds a spatial lag of the explanatory variables to generate a second set of coefficients (β_2) that capture spatial dependence in the independent variables:

$$y = \rho Wy + X\beta_1 + WX\beta_2 + u \tag{4}$$

while the spatial Durbin error model (SDEM) specification includes the spatial lag of the explanatory variables and accounts for spatial dependence in the error term:

$$y = X\beta_1 + WX\beta_2 + \lambda Wu + \varepsilon \tag{5}$$

In practice, without strong theoretical justification, it is often difficult to determine through specific diagnostic tests exactly which model specification should be used (Anselin and Rey 2014; LeSage 2014). Given that fact—and this article's goal to test these traditional model specifications with random forest regressors—the approach used here is to compare each specification based on its prediction accuracy, MAPE, in two testing contexts (described in the Results below). All spatial econometric models were estimated using the spdep package in R v.3.5.1.

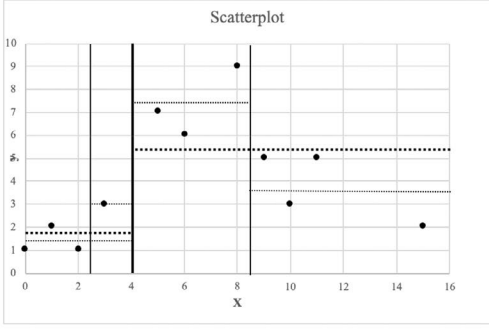
Random forest

A "random forest" is an aggregation of the results of many individual decision trees, so it is first necessary to describe the Classification and Decision Tree (CART) training algorithm. Decision trees can be used for either *classification* or *regression* problems; these produce category probabilities or average predicted continuous values, respectively (Géron 2017; Krzywinski and Altman 2017). This article is interested in regression results, as the dependent variables of interest are continuous values, so the methods described here will focus on regression trees. CART is a nonlinear function that iteratively splits the training data into subsets using the threshold criterion that produces subsets with the lowest weighted average mean squared error (MSE) between predicted and observed y values (Krzywinski and Altman 2017). A simple example using randomly generated data is shown below in Fig. 2.

The first naïve predicted value for $\bar{y}$ in the CART algorithm is simply the average of all y values (4, in this example). By iteratively dividing the data, the algorithm then determines that $X \leq 4$ provides the smallest weighted average of MSE of both subsets (3.289). The new $\ddot{y}$ values predicted by the decision tree are now the average y values of each of these subsets (marked by the thick dotted lines on the scatterplot)—5.29 for observations with X values above 4 and 1.75 for observations with X values less than or equal to 4. The algorithm now iterates across both of these two subsets, finding that 2.5 and 8.5 most reduce the total weighted average of MSE of all subsets (1.098). The new predicted $\ddot{y}$ values are again the average y values of each of the four subsets (marked by the thin dotted lines on the scatterplot)—1.33 for observations with X values from 0 to 2.5, 3 for observations with X values from 2.5 to 4, 7.33 for observations with X values from 4 to 8.5, and (back down to) 3.75 for observations with X values above 8.5. The shape of this resulting "curve" now much more accurately follows the pattern of the data. The CART algorithm ends when it reaches a prespecified depth (in this case, 2).

The general idea of the random forest—and all "ensemble" methods—is based on the Law of Large Numbers: the more predictions that are made, the more likely they are to average to the true expected value, since fluctuations in positive and negative error for predicted values will tend to cancel one another out with a large enough number of trials (Breiman 2001a). A random

X	y	ȳ1	MSE1	Sum (MSE1)	ȳ2	MSE2	Sum (MSE2)	ȳ3	MSE3	Sum (MSE3)
0	1	4	0.8182		1.7500	0.1406		1.3333	0.0370	
1	2	4	0.3636		1.7500	0.0156	0.6875	1.3333	0.1481	0.2222
2	1	4	0.8182		1.7500	0.1406		1.3333	0.0370	
3	3	4	0.0909		1.7500	0.3906		3	0	0
5	7	4	0.8182		5.2857	0.4198		7.3333	0.0370	
6	6	4	0.3636	6.1818	5.2857	0.0729		7.3333	0.5926	1.5556
8	9	4	2.2727		5.2857	1.9708		7.3333	0.9259	
9	5	4	0.0909		5.2857	0.0117	4.7755	3.7500	0.3906	
10	3	4	0.0909		5.2857	0.7464		3.7500	0.1406	1.6875
11	5	4	0.0909		5.2857	0.0117		3.7500	0.3906	
15	2	4	0.3636		5.2857	1.5423		3.7500	0.7656	
Weighted average MSE of subsets				*6.182*			*3.289*			*1.098*



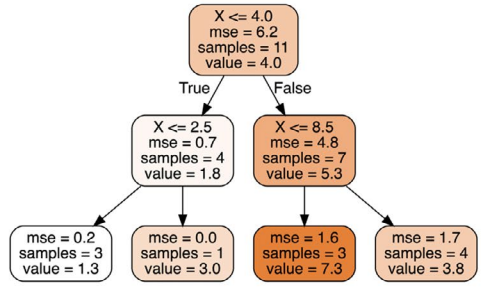


Figure 2. Example of the Classification and Decision Tree (CART) algorithm using a random data set.

forest is an ensemble of (e.g., 1,000) decision trees that are trained on random subsets of the data with replacement, which is known "bootstrap aggregating" or "bagging" (Breiman 1996). The predicted values from these individual trees are then averaged. Of course, better predictions from an ensemble will result when individual predictors are as independent as possible, so to decrease correlation between individual trees, the random forest classifier only allows each tree to use a random subset of the explanatory variables, which produces more diversity among trees (Breiman 2001a; Géron 2017).

For this analysis, four different RF specifications were used (each calculated with 1,500 decision trees with the default setting for tree depth[12] using the *sci-kit learn* function "RandomForestRegressor" in Python 3): a baseline model with the same covariates as used in the spatial econometric models above ("RF"), a model built to approximate the SAR model by including the spatially lagged dependent variable ("RFSAR"), a model including spatial lags of both the dependent and all independent variables meant to approximate the SDM ("RFSDM"), and finally, a model including only the spatial lags of the independent variables (similar to the spatial lag of *X* model) ("RFSLX"). Evaluating RF models with various spatial parameters provides a first look at understanding (and disentangling) the role of space in this predictive method.

Results

The baseline data used to train all nine models come from 2010, with LOGE10SEB as the dependent variable and six 2010 covariates (shown in Table 2). Again, the approach used in this article is to use the estimated relationships between these variables to predict observed expected employment density around the new Expo Line in the post-construction period. Table 2 displays the significant (at $P \leq .05$) regression coefficients for the various SE models and the "permutation importances" for the RF models for these baseline models. Permutation importances are

Table 2. Coefficients and Permutation Importances of Various Spatial Econometric and Random Forest Models

		Coefficients					Permutation importances			
	Independent variables	OLS	SAR	SEM	SDM	SDEM	RF	RFSAR	RFSDM	RFSLX
X	LOGE04SEB	**0.524**	**0.429**	**0.485**	**0.464**	**0.467**	53.24%	34.14%	37.29%	39.60%
	LOGR10SEB	**0.113**	**0.083**	**0.043**	**0.021**	**0.031**	14.93%	7.28%	3.89%	7.81%
	E_DIV10	**0.772**	**0.688**	**0.701**	**0.690**	**0.702**	43.66%	35.43%	34.25%	36.71%
	R_YPCT10	**−0.208**	**−0.159**	**−0.176**	**−0.183**	**−0.192**	2.14%	1.11%	0.79%	0.93%
	R_IPCT10	**−0.598**	**−0.439**	**−0.322**	**−0.264**	**−0.305**	5.63%	2.12%	1.05%	2.01%
	TRANS10	**0.157**	**0.067**	**0.139**	**0.004**	0.008	0.09%	0.02%	−0.0002%	0.002%
Lags of *X*	LOGE04SEB				**−0.174**	**0.088**			3.21%	3.34%
	LOGR10SEB				**0.120**	**0.204**			0.44%	1.20%
	E_DIV10				**−0.124**	**0.321**			0.56%	0.93%
	R_YPCT10				**−0.144**	**−0.283**			0.14%	0.23%
	R_IPCT10				**−0.408**	**−0.867**			0.33%	2.06%
	TRANS10				**0.041**	**0.090**			0.01%	0.06%
Lag of *y*	LOGE10SEBL							9.03%	18.98%	
Spatial Variables	Rho		**0.348**		**0.482**					
	Lambda			**0.588**		**0.482**				
AIC		1,02,330	95,040	94,873	92,826	92,936				

Note: Regression coefficients in bold are significant at $P \leq .05$ level; insignificant coefficients are not bolded.

calculated by randomly shuffling the values of a given variable in the "out-of-bag" (OOB) sample (the data that were not used during tree training); the average decrease in prediction accuracy of the OOB sample using the shuffled (permuted) variable across all trees in the forest is that variable's permutation importance (or, similarly, the percentage *increase* in prediction accuracy obtained by including that variable in the model) (Breiman and Cutler 2004; Parr et al. 2018). Permutation importances are not directly provided by *sci-kit learn* but are less influenced by correlated variables than the standard impurity importances reported by default (Strobl et al. 2007).

As Table 2 shows, the results of SE models are fairly robust across the various specifications. The ρ and λ coefficients are significant across the various models, suggesting that spatial autocorrelation is present in the data and thus needs to be controlled for. In terms of the application's substantive question, the TRANS10 variable is significant in several specifications, suggesting initially that transit proximity may have a significant positive relationship with expected employment density, although the variable is not significant in the Error Durbin model. The effects of R_YPCT10 and R_IPCT10 on expected employment density are negative, and the Durbin model has lowest the Akaike information criterion (AIC) value, suggesting—under this traditional model selection criteria—this it is the preferred specification.

The RF model results are also relatively robust across specifications; perhaps unsurprisingly, LOGE04SEB displays a high importance across all specifications (expected employment density in 2004 would logically play an important role in predicting 2010 expected employment density), but more interesting from a spatial perspective is that the spatial lag of *y* variable (LOGE10SEBL) displays a high importance in the two models that include it. Also, TRANS10 has a very low importance in all RF models, even displaying a *negative* importance in the RFSDM model, which implies that its inclusion actually makes that model's accuracy slightly worse. These findings are further augmented by examining partial dependence plots (PDP) for the LOGE10SEBL and TRANS10 variables in the RFSDM model, shown in Fig. 3. PDP show the marginal effect that a variable of interest (x) has on the response variable (y),[13] thus providing information similar in concept to a regression coefficient, although, unlike in linear regression, the relationships displayed for RF models can be nonlinear (Molnar 2018). As Fig. 3 shows, the average marginal effect of the spatial lag (LOGE10SEBL, x-axis) on the dependent variable (LOGE10SEB, y-axis) is positive and roughly linear, although the dependent variable does not appear to increase much in observations with spatial lag values above −3. Even more interesting is the result for TRANS10, which indicates a very flat marginal effect across all observations and dipping slightly negative for observations in which TRANS10 = 1. In this case, the PDP mirrors the low permutation importance and coefficient size for this variable in the RFSDM and SDM specifications, respectively.

Table 3 displays the comparison in model accuracy based on MAPE (calculated by taking 1—(the average across all values of the absolute value of (predicted y—observed y)/observed y), which expresses the error value as a percentage of the observed value) for the baseline 2010 model and the post-construction model. In the first scenario, each model is trained on the full data set[14] and tested by comparing predicted LOGE10SEB values to observed LOGE10SEB values in a random subset of 20% of the data. In the second scenario, the models are trained in a similar way (with TRANS = 1 updated for Expo Line-proximate blocks) but tested by comparing predicted LOGE10SEB values to observed post-construction LOGEPSSEB values in blocks around the newly constructed Expo Line.

Interestingly, the RF models (as a set) provide a slight increase in predictive accuracy in both scenarios compared to the traditional SE models, which supports previous research that RF performs best for predicting certain social-geographical variables such as parking violations

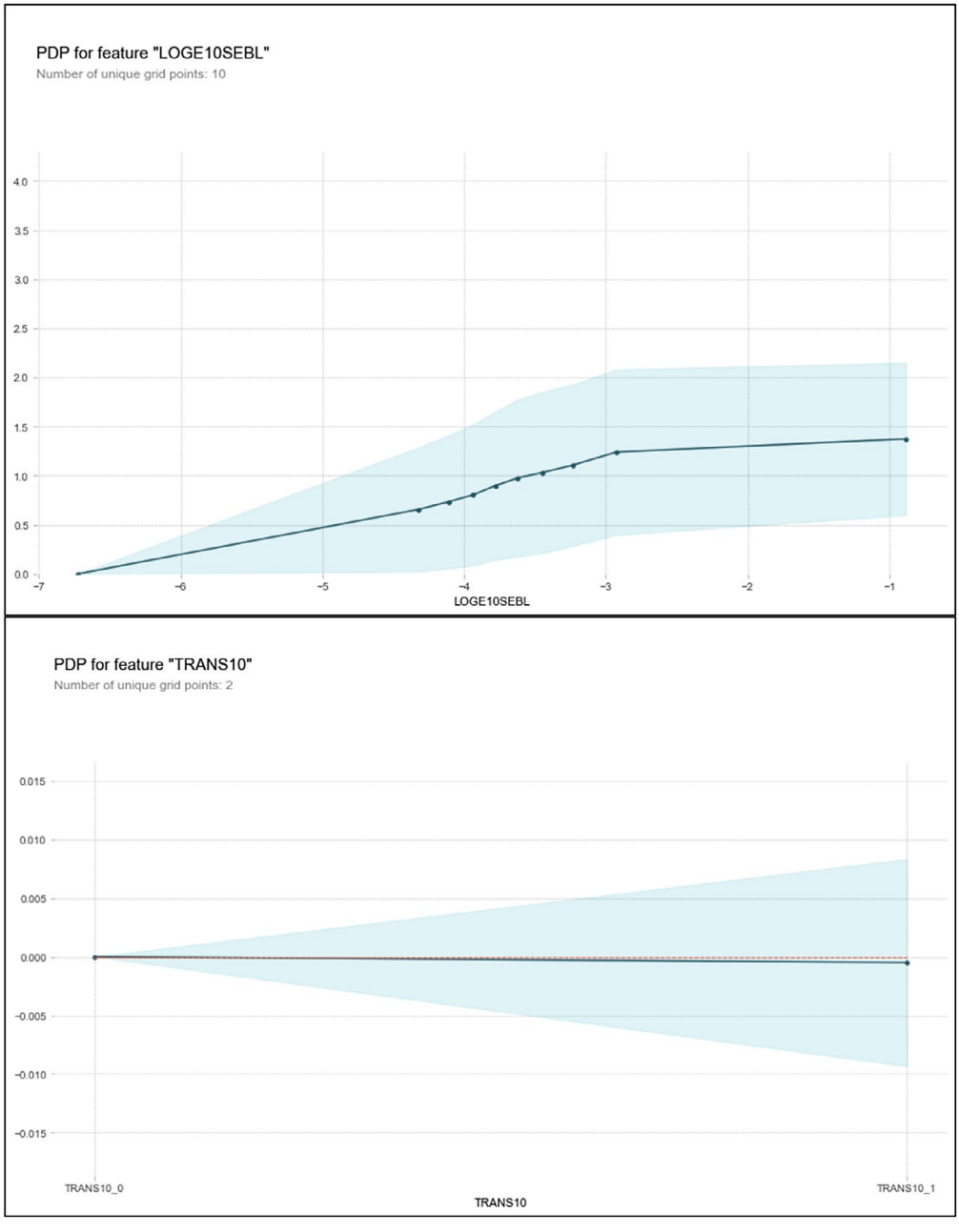


Figure 3. Partial dependence plots (PDP) showing the average marginal effect of LOGE10SEBL (top) and TRANS10 (bottom) on LOGE10SEB in the RFSDM model.

(Gao et al. 2019). Within the SE models, the OLS model performs best in Scenario 1,[15] while the SDM performs best in Scenario 2, as the AIC suggested. For the RF models, the RFSDM specification demonstrates the highest prediction accuracy out of all specifications in both scenarios. And while predictive accuracy declines across the board for the new construction scenario—which makes sense, given the longer time lag (and thus possible introduction of error) between

Table 3. Comparison of Model Accuracy for Various Spatial Econometric and Random Forest Models

		Scenario 1	Scenario 2
Training *y*		LOGE10SEB	LOGE10SEB
Testing *y*		LOGE10SEB	LOGEPSSEB
Training Set		Random 80%	Full data
Testing Set		Random 20%	EXPO = 1 (w/TRANS variable updated = 1)
Accuracy (1—Mean Absolute Percentage Error)	OLS	88.48%	80.72%
	SAR	85.94%	81.74%
	SEM	88.22%	81.72%
	SDM	85.45%	81.88%
	SDEM	86.02%	81.85%
	RF	90.93%	84.37%
	RFSAR	91.41%	84.58%
	RFSDM	**91.75%**	**84.61%**
	RFSLX	91.38%	84.49%

Selected (best-performance) model results highlighted in bold

predicted and observed values (in the case of Phase II blocks, a lag from 2010 to 2017)—the RFSDM model still provides an average prediction within 85% of the observed values.

While the RF models modestly outperform their SE counterparts in each case, it is also important to better understand the spatial characteristics driving RF versus SE performance. To do this, correlations between the explanatory variables used in the models and "MAPE_DIF" (SDM absolute percentage error—RFSDM[16] absolute percentage error for each block for the post-construction model) are calculated (shown in Table 4). Thus, negative values of MAPE_DIF indicate blocks in which the RFSDM outperformed SDM; positive values indicate blocks in which SDM outperformed RFSDM. Along the same lines, negative correlations indicate that RFSDM tends to outperform SDM in blocks with larger values of LOGR10SEB and R_IPCT10; conversely, positive correlations indicate that SDM tends to outperform RFSDM in blocks with larger values of the two dependent variables (LOGEPSSEB and LOGE10SEB).

As Table 4 shows, the correlations are generally quite low, but it appears that blocks in which RF outperforms SE tend to have larger expected employment densities (LOGEPSSEB and LOGE10SEB), and blocks in which SE outperforms RF tend to have larger expected residential densities and larger percentages of residents with high incomes (LOGR10SEB and R_IPCT10).[17] It is also possible that predictive performance follows some kind of spatial pattern, so Fig. 4 shows the mapped MAPE_DIF values. While the pattern appears to be quite heterogeneous (as the correlations also suggest), this method is useful as a check to help more completely understand the role of space in the predictive performance of RF models.

Discussion and conclusion

This article evaluates the use of spatially explicit machine learning models in predicting employment around a newly constructed transit line in Los Angeles in order to better understand how these models make predictions and whether the inclusion of spatial variables improves their performance. The results of this analysis show that random forest models modestly outperform

Table 4. Correlations Between Explanatory Variables and MAPE_DIF

Correlation Variable	MAPE_DIF
LOGEPSSEB	0.18
LOGE10SEB	0.17
LOGR10SEB	*–0.07*
LOGE04SEB	0.04
E_DIV10	0.02
R_YPCT10	–0.01
R_IPCT10	*–0.06*
LOGEPSEBL	0.02
LOGE10SEBL	0.03
LOGR10SEBL	–0.02
LOGE04SEBL	0.00
E_DIV10L	–0.02
R_YPCT10L	0.02
R_IPCT10L	–0.05

traditional spatial econometric models in each test case (although the differences are not large enough to settle the question conclusively). When testing predictions on 2010 expected employment density based on 2010 relationships, RF models generally provide around a 2% increase in accuracy. Interestingly, when testing predictions based on post-construction expected employment densities in the context of the newly constructed Expo Line 3–7 years after the training year, RF models perform even better, averaging a 3% increase in accuracy compared to SE models—the best-fit spatial random forest model demonstrates 84.61% accuracy in predicting post-construction employment density around newly built transit stations, compared to 81.88% for the best-fit spatial econometric model. While these differences are small, this suggests that RF models could be an even better choice than traditional models when prediction uncertainty and error are larger. RF models also perform particularly well (compared to SE models) in blocks with larger expected employment densities, although the spatial pattern of performance on the block level is quite heterogeneous, making it hard to draw any general conclusions in this context.

This article also puts forward a unique strategy for building spatially explicit predictive machine learning models by using spatially lagged variables in the RF specification to mirror the specification of various traditional SE models. The inclusion of spatial lags slightly improves predictive performance for these models (relative to the standard RF model), and spatial lags of the dependent variable (in particular) show relatively high permutation importances, relating to a 9% to 19% increase in model accuracy based on the inclusion of the spatial lag of y, depending on model type. While additional work on predictive machine learning models in a spatial context is certainly needed, this article provides an important first step in testing their predictive effectiveness—and the effectiveness (albeit small) of including spatial variables in these models to improve their performance.

Given this article's development of a framework for including spatially explicit data in random forest models, spatial data scientists and geographers may also want to begin to consider using the RFSDM specification for more explanatory-focused research questions. While the usefulness of RF models in the explanatory context appears to depend on the specific question at hand—and additional work needs to be done to continue to open the "black box" of RF methods, such as the use of "causal" trees (Athey and Imbens 2016)—this article's application

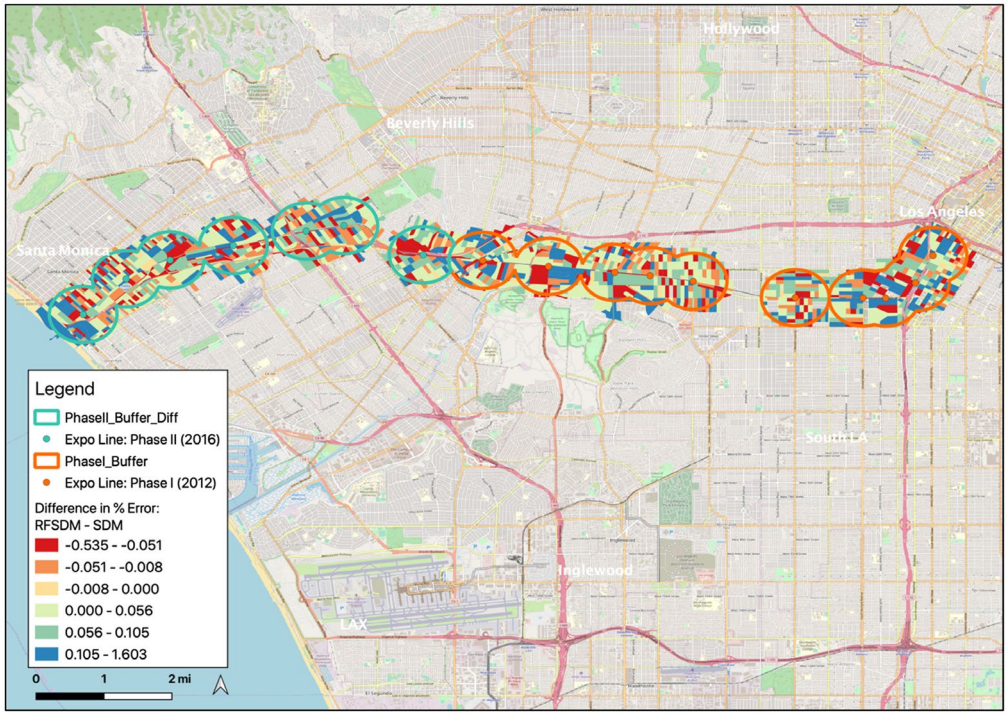


Figure 4. Map showing difference in predictive performance between RFSDM and SDM models.

demonstrates that features like permutation importances and PDP can be used to interpret the explanatory relationships produced by random forest models.

Notes

1 Credit (2019) and Giuliano and Agarwal (2017) provide a more thorough discussion of the theoretical basis for the expected relationship between various economic development indicators and increased transportation accessibility (both generally and in the specific context of public transit). These articles also review the (large) existing empirical literature on economic development and transit; a number of studies have used employment or employment density as the dependent variable of interest (e.g., Knight and Trygg 1977; Green and James 1993; Bollinger and Ihlanfeldt 1997; Cervero and Landis 1997), although most recent studies examine property values or new business creation.

2 While the baseline comparison made in this article is to test identical parsimonious specifications, one of the advantages of the random forest method in the prediction context is that these models are not constrained by the assumption of linearity. This means that the prediction accuracy shown here is only a baseline and could likely be improved substantially (compared to what is statistically feasible in spatial econometric models) by using additionally more complicated specifications without the constraint of the linearity assumption or the danger of biased coefficient estimates, etc.

3 Often for good reason; as Rey (2019) points out, these predictive methods are often combined with a mindset of "Code Hubris" by data scientists who apply them to a wide array of geographical and social research problems with little underlying knowledge of geographic or social science theory.

4 For the SE models, computations were made using the *lagsarlm* function in R, with a 12 nearest-neighbor spatial weights matrix for both the full dataset and the 20% testing sample, zeros assigned to lagged values without neighbors, and lagged variables calculated independently between in-sample and out-of-sample units (Bivand 2018).

5 To clarify, the variables included in the training set for Scenario 1 and Scenario 2—including the spatial lag of the dependent variable in 2010, LOGE10SEBL—are exactly the same. Thus the SAR, SDM, RFSAR, and RFSDM Scenario 2 specifications *do not* employ spatial lags of the post-construction dependent variable; instead, they are trained on the 2010 relationships and compared to post-construction values (only in the treatment area) using MAPE.

6 The variable "POSTETOT" was constructed by taking the post-construction employment values for blocks intersecting an 800 m buffer around Expo Line stations. Around the stations constructed in Phase I (2012), 2013 employment values are used; around the stations constructed in Phase II (2016), 2017 employment values are used. Thus the POSTETOT variable represents amalgamated post-construction employment.

7 Employment by NAICS code was used to create a measure of employment diversity by calculating a Herfindahl index based on 2-digit NAICS categories (and inverting that value so that larger values indicate more diversity): "E_DIV10".

8 For more information on the calculation of the SEB smoother, see Clayton and Kaldor (1987) and Anselin, Kim, and Syabri (2004). All SEB calculations used in this paper were performed in GeoDa v.1.14.0.

9 However, in very large samples (which we have in this case), the assumption of normality is less important (Lumley et al. 2002).

10 Using the SEB smoother, 87,227 out of 100,319 "urban"-designated blocks in LA County are nonzero observations rather than 63,885 out of 100,319 using the raw employment density variable.

11 While researchers have rightfully called for a more rigorous process for justifying the choice of weights matrices (LeSage and Pace 2010), in this case the kinds of Bayesian model comparison approaches most commonly used to select optimal model-weights matrix combinations (LeSage 2014, 2015) cannot be used, since the goal is to compare spatial econometric models to random forest models using predictive accuracy as the common metric. To ensure consistency across these very different kinds of models, a single weights matrix was chosen.

12 From the *sci-kit learn* documentation on default cut-off depth: "nodes are expanded until all leaves are pure or until all leaves contain less than min_samples_split samples," which in this case is 2.

13 From Molnar (2018): "The computation of partial dependence plots is intuitive: The partial dependence function at a particular feature value represents the average prediction if we force all data points to assume that feature value…If the feature for which you computed the PDP is not correlated with the other features, then the PDPs perfectly represent how the feature influences the prediction on average."

14 SE models are trained using the full data set and RF models are trained on a random 80% of the data. The use of an 80% random sample rather than the full data set is due to standard practice in RF model constriction to avoid overfitting, a concern that is not present for SE models (Géron 2017).

15 While it is somewhat surprising that the baseline OLS performed best in this case, it may be due to the fact that performing out-of-sample prediction for spatial models is not so straightforwardβsince the spatial weights matrix is directly inserted into the estimation of the model, predicting to the out-of-sample context requires the use of a spatial weights matrix for the random subset of blocks in the out-of-sample group, which necessarily creates gaps in the original fabric of blocks and likely entails a lower predictive value for the spatial parameters (since the randomization of the testing set necessarily removes the spatial configuration that might lend added value to the spatial parameters) (Bivand 2018). An alternative version of Scenario 1, with the SE models trained on the full dataset and predicted to a random 20% subset yielded similar (slightly larger) overall MAPE values, with the SDM displaying the highest accuracy out of all SE models. All SE models still displayed lower accuracy than all RF models in this case. Detailed results are available from the author at request.

16 The Durbin versions of SE and RF models were chosen because they demonstrated the highest prediction accuracy of all models in Scenario 2.

17 Of course, the reverse cases could be true. Also, since MAPE_DIF is a continuous variable, the relationships are not necessarily always split along positive/negative lines.

APPENDIX

Table A1. Variable Names, Descriptions, Sources, and Descriptive Statistics for Variables Used in Models

Variable	Description	Source	Min	25%	Median	Mean	75%	Max
E_TOT10	2010 employment (workplace)	LODES	0	0	3	45.41	15	53,611
E_DENS10	2010 employment density (workplace)	LODES	0	0	0.0001433	0.0017316	0.0008915	1.6651121
LOGE10SEB	Log10(expected 2010 employment density)	Spatial Empirical Bayes adjustment (using 12 nearest neighbors)	−8.4074	−4.2486	−3.751	−3.6633	−3.0494	0.2214
LOGE10SEBL	Spatial lag of expected 2010 employment density	12-nearest neighbors spatial weights matrix	−6.7252	−4.0651	−3.7035	−3.6607	−3.2928	−0.8845
E_TOT04	2004 employment (workplace)	LODES	0	0	1	43.15	12	52,456
E_DENS04	2004 employment density (workplace)	LODES	0	0	0.000037	0.0016421	0.0007339	1.9056595

(Continues)

Table A1. (Continued)

Variable	Description	Source	Min	25%	Median	Mean	75%	Max
LOGE04SEB	Log10(expected 2004 employment density)	Spatial Empirical Bayes adjustment (using 12 nearest neighbors)	−8.848	−4.761	−4.068	−3.945	−3.128	0.28
LOGE04SEBL	Spatial lag of expected 2004 employment density	12-nearest neighbors spatial weights matrix	−6.6643	−4.4442	−4.0139	−3.9389	−3.475	−0.9171
R_TOT10	2010 employment (residence)	LODES	0	7	27	43.45	55	1,597
R_DENS10	2010 employment density (residence)	LODES	0	0.000575	0.001425	0.001814	0.002374	0.232652
LOGR10SEB	Log10(expected 2010 employed residential density)	Spatial Empirical Bayes adjustment (using 12 nearest neighbors)	−9.288	−3.168	−2.838	−3.042	−2.637	−0.664

(Continues)

Table A1. (Continued)

Variable	Description	Source	Min	25%	Median	Mean	75%	Max
LOGR10SEBL	Spatial lag of expected 2010 employment density (residence)	12-nearest neighbors spatial weights matrix	−6.722	−3.14	−2.883	−3.005	−2.721	−1.847
POSTETOT	For Phase I blocks—2013 employment (workplace); for Phase II blocks—2017 employment (workplace)	LODES	0	2	10	83.86	60	3,275
E_DPOST	POSTETOT density	LODES	0	0.00008	0.00055	0.00493	0.00349	0.27118
LOGEPSSEB	Log10(expected POSTETOT density)	Spatial Empirical Bayes adjustment (using 12 nearest neighbors)	−7.370	−4.580	−3.640	−3.630	−2.690	−0.570

(Continues)

Table A1. (Continued)

Variable	Description	Source	Min	25%	Median	Mean	75%	Max
LOGEPSEBL	Spatial lag of expected POSTETOT density	12-nearest neighbors spatial weights matrix	−6.310	−4.540	−3.870	−3.890	−3.250	−1.400
E_DIV10	2010 employment diversity (workplace)	1—(Herfindahl Index for employment by 2-digit NAICS categories)	0	0	0	0.2059	0.449	0.9002
E_DIV10L	Spatial lag of 2010 employment diversity (workplace)	12-nearest neighbors spatial weights matrix	0	0.09404	0.17414	0.19961	0.28089	0.77539
R_YPCT10	2010 percent employed residents aged 29 or younger	LODES	0	0.1136	0.2059	0.1995	0.2734	1
R_YPCT10L	Spatial lag of 2010 percent employed residents aged 29 or younger	12-nearest neighbors spatial weights matrix	0	0.1629	0.2043	0.2011	0.2442	0.4663

(Continues)

Table A1. (Continued)

Variable	Description	Source	Min	25%	Median	Mean	75%	Max
R_IPCT10	2010 percent employed residents with earnings > $3,333/month	LODES	0	0.2043	0.3492	0.3407	0.4848	1
R_IPCT10L	Spatial lag of 2010 percent employed residents with earnings > $3,333/month	12-nearest neighbors spatial weights matrix	0	0.2446	0.3411	0.3423	0.4393	0.7666

References

Agostini, C. A., and G. Palmucci. (2008). "The Anticipated Capitalization Effect of a New Metro Line on Housing Prices." *Fiscal Studies* 29, 233–56.

Alejandro, Y., and L. Palafox. (2019). "Gentrification Prediction Using Machine Learning." In Advances in Soft Computing. MICAI 2019. Lecture Notes in Computer Science, vol 11835, edited by L. Martínez-Villaseñor, I. Batyrshin, and A. Marín-Hernández. Cham: Springer. https://doi.org/10.1007/978-3-030-33749-0_16.

Allen, C., M. H. Tsou, A. Aslam, A. Nagei, and J. M. Gawron. (2016). "Applying GIS and Machine Learning Methods to Twitter Data for Multiscale Surveillance of Influenza." *PLoS One* 11(7), e0157734. https://doi.org/10.1371/journal.pone.0157734.

American Public Transportation Association (APTA). (2019). "Transit Ridership Report: Third Quarter 2019." American Public Transportation Association. Retrieved from: https://www.apta.com/wp-content/uploads/2019-Q3-Ridership-APTA.pdf.

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

Anselin, L., Y. W. Kim, and I. Syabri. (2004). "Web-Based Analytical Tools for the Exploration of Spatial Data." *Journal of Geographical Systems* 6, 197–218.

Anselin, L., and S. Rey. (2014). Modern Spatial Econometrics in Practice. Chicago, IL: GeoDa Press LLC.

Athey, S., and G. Imbens. (2016). "Recursive Partitioning for Heterogeneous Causal Effects." *Proceedings of the National Academy of Sciences* 113(27), 7353–60.

Bivand, R. (2018). predict.sarlm: Prediction For Spatial Simultaneous Autoregressive Linear Model Objects. Retrieved from: https://www.rdocumentation.org/packages/spdep/versions/0.7-7/topics/predict.sarlm.

Bollinger, C. R., and K. R. Ihlanfeldt. (1997). "The Impact of Rapid Rail Transit on Economic Development: The Case of Atlanta's MARTA." *Journal of Urban Economics* 42, 179–204.

Breiman, L. (1996). "Bagging Predictors." *Machine Learning* 24(2), 123–40.

Breiman, L. (2001a). "Random Forests." *Machine Learning* 45, 5–32.

Breiman, L. (2001b). "Statistical Modeling: The Two Cultures." *Statistical Science* 16(3), 199–231.

Breiman, L., and A. Cutler. (2004). "Random Forests." Retrieved from: https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm-varimp.

Cervero, R. (1984). "Light Rail Transit and Urban Development." *Journal of the American Planning Association* 50(2), 133–47.

Cervero, R. (1994). "Rail Transit and Joint Development: Land Market Impacts in Washington, D.C. and Atlanta." *Journal of the American Planning Association* 60(1), 83–94.

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), 121–38.

Cervero, R., and J. Landis. (1997). "Twenty Years of the Bay Area Rapid Transit System: Land Use and Development Impacts." *Transportation Research Part A* 41(4), 309–33.

Chatman, D. G., R. B. Noland, and N. J. Klein. (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. https://doi.org/10.3141/2598-01.

Clayton, D., and J. Kaldor. (1987). "Empirical Bayes Estimates of Age-Standardized Relative Risks for Use in Disease Mapping." *Biometrics* 43, 671–81.

Credit, K. (2018). "Transit-oriented economic development: The impact of light rail on new business starts in the Phoenix, AZ Region." *Urban Studies* 55(13), 2838–62. https://doi.org/10.1177/0042098017724119.

Credit, K. (2019). "Transitive Properties: A Spatial Econometric Analysis of New Business Creation Around Transit." *Spatial Economic Analysis* 14(1), 26–52.

Damm, D., S. R. Lerman, E. Lerner-Lam, and J. Young. (1980). "Response of Urban Real Estate Values in Anticipation of the Washington Metro." *Journal of Transport Economics and Policy* 14(3), 315–36.

Fan, C., Z. Cui, and X. Zhong. (2018). "House Prices Prediction with Machine Learning Algorithms." Proceedings of the 2018 10th International Conference on Machine Learning and Computing—ICMLC 2018. https://doi.org/10.1145/3195106.3195133.

Feng, Y., Z. Cai, X. Tong, J. Wang, C. Gao, S. Chen, and Z. Lei. (2018). "Urban Growth Modeling and Future Scenario Projection Using Cellular Automata (CA) Models and the R Package Optimix." *International Journal of Geo-Information* 7. https://doi.org/10.3390/ijgi7100387.1–20.

Gao, S., M. Li, Y. Liang, J. Marks, Y. Kang, and M. Li. (2019). "Predicting the Spatiotemporal Legality of On-Street Parking using Open Data and Machine Learning." *Annals of GIS* 25(4), 299–312. https://doi.org/10.1080/19475683.2019.1679882.

Georganos, S., T. Grippa, N. Gadiaga, C. Linard, M. Lennert, S. Vanhuysse, N. Mboga, E. Wolff, and S. Kalogirou. (2019). "Geographical Random Forests: A Spatial Extension to the Random Forest Algorithm to Address Spatial Heterogeneity in Remote Sensing and Population Modeling." *Geocarto International*. https://doi.org/10.1080/10106049.2019.1595177.1–16.

Géron, A. (2017). Hands-On Machine Learning with Scikit-Learn & TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. Boston, MA: O'Reilly.

Giuliano, G., and A. Agarwal. (2017). "Land Use Impacts of Transportation Investments." In The Geography of Urban Transportation, 4th ed., edited by S. Hanson and G. Giuliano. New York, NY: Guilford Press.218–245.

Golub, A., S. Guhathakurta, and B. Sollapuram. (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), 415–29. https://doi.org/10.1177/0739456X12455523.

Graham, M. R., M. J. Kutzbach, and B. McKenzie. (2014). "Design Comparison of LODES and ACS Commuting Data Products". Working Papers, Center for Economic Studies, U.S. Census Bureau, 14–38. Retrieved from: https://ideas.repec.org/p/cen/wpaper/14-38.html.

Green, R. D., and D. M. James. (1993). Rail Transit Station Area Development: Small Area Modeling in Washington, DC. Armonk, NY: M.E. Sharpe.

Grekousis, G. (2019). "Artificial Neural Networks and Deep Learning in Urban Geography: A Systematic Review and Meta-Analysis." *Computers, Environment and Urban Systems* 74, 244–56.

Hengl, T., M. Nussbaum, M. N. Wright, G. B. M. Heuvelink, and B. Gräler. (2018). "Random Forest as a Generic Framework for Predictive Modeling of Spatial and Spatio-Temporal Variables." *PeerJ* 6, e5518. https://doi.org/10.7717/peerj.5518.

Hess, D. B., and T. M. Almeida. (2007). "Impact of Proximity to Light Rail Rapid Transit on Station-area Property Values in Buffalo, New York." *Urban Studies* 44(5–6), 1041–68.

Janowicz, K., S. Gao, G. McKenzie, Y. Hu, and B. Bhaduri. (2019). "GeoAI: Spatially Explicit Artificial Intelligence Techniques for Geographic Knowledge Discovery and Beyond." *International Journal of Geographical Information Science* 34(4), 625–36. https://doi.org/10.1080/13658816.2019.1684500.

Knaap, G., J. C. Ding, and L. D. Hopkins. (2001). "Do Plans Matter? The Effects of Light Rail Plans on Land Values in Station Areas." *Journal of Planning Education and Research* 21(1), 32–9.

Knight, R. L., and L. L. Trygg. (1977). "Evidence of Land Use Impacts of Rapid Transit Systems." *Transportation* 6(3), 231–47.

Knorr, D. (2019). "Using Machine Learning to Identify and Predict Gentrification in Nashville, Tennessee." Thesis: Master of Science in Earth and Environmental Science at Vanderbilt Unievrsity. Retrieved from: https://ir.vanderbilt.edu/bitstream/handle/1803/13285/07242019_Dknorr_Thesis_Final2.pdf?sequence=1&isAllowed=y.

Krzywinski, M., and N. Altman. (2017). "Classification and Regression Trees." *Nature Methods* 14, 757–8.

Landis, J., S. Guhathakurta, and M. Zhang. (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.

LeSage, J. P. (2014). "Spatial Econometric Panel Data Model Specification: A Bayesian Approach." *Spatial Statistics* 9, 122–45.

LeSage, J. P., and R. K. Pace. (2009). Introduction to Spatial Econometrics. New York, NY: CRC Press.

LeSage, J.P. & Pace, R.K.(2010) The Biggest Myth in Spatial Econometrics.(December 1, 2010) *Available at SSRN:*, https://ssrn.com/abstract=1725503.

LeSage, J.P. (2015) Software for Bayesian cross section and panel spatial model comparison. *Journal of Geographical Systems*, 17 297–310. https://doi.org/10.1007/s10109-015-0217-3.

Lumley, T., P. Diehr, S. Emerson, and L. Chen. (2002). "The Importance of the Normality Assumption in Large Public Health Data Sets." *Annual Review of Public Health* 23, 151–69.

Mohammad, S., D. Graham, P. Melo, and R. Anderson. (2013). "A Meta-Analysis of the Impact of Rail Projects on Land and Property Values." *Transportation Research Part A* 50, 158–70.

Molnar, C. (2018). Interpretative Machine Learning: A Guide for Making Black Box Models Explainable. Retrieved from: https://christophm.github.io/interpretable-ml-book/pdp.html.

Mu, J., F. Wu, and A. Zhang. (2014). "Housing Value Forecasting Based on Machine Learning Methods." *Abstract and Applied Analysis* 2014, 1–7. https://doi.org/10.1155/2014/648047.

Openshaw, S., and C. Openshaw. (1997). Artificial Intelligence in Geography. New York, NY: John Wiley & Sons Inc.

Parr, T., K. Turgutlu, C. Csiszar, and J. Howard. (2018). "Beware Default Random Forest Importances." Retrieved from: https://explained.ai/rf-importance/.

Phan, T. D. (2018). "Housing Price Prediction Using Machine Learning Algorithms: The Case of Melbourne City, Australia." International Conference on Machine Learning and Data Engineering (ICMLDE) 2018. https://doi.org/10.1109/icmlde.2018.00017.

Prateek, J. (2017). Artificial Intelligence with Python: A Comprehensive Guide to Building Intelligent Apps for Python Beginners and Developers. Birmingham, UK: Packt Publishing.

Rey, S. (2019). "Geographical Analysis: Reflections of a Recovering Editor." *Geographical Analysis* 1–9. https://doi.org/10.1111/gean.12193.

Seo, K., A. Golub, and M. Kuby. (2014). "Combined Impacts of Highways and Light Rail Transit on Residential Property Values: A Spatial Hedonic Price Model for Phoenix, Arizona." *Journal of Transport Geography* 41, 53–62.

Singleton, A., and D. Arribas-Bel. (2019). "Geographic Data Science." *Geographical Analysis* 1–15. https://doi.org/10.1111/gean.12194.

Strobl, C., A. L. Boulesteix, A. Zeileis, and T. Hothorn. (2007). "Bias in Random Forest Variable Importance Measures: Illustrations, Courses and a Solution." *BMC Bioinformatics* 8(25), 1–21.

Tobler, W. (1970). "A Computer Movie Simulating Urban Growth in the Detroit Region." *Economic Geography* 46, 234–40.

Truong, Q., Nguyen, M., Dang, H., and Mei, B. (2020). "Housing Price Prediction via Improved Machine Learning Techniques. 2019 International Conference on Identification, Information and Knowledge in the Internet of Things (IIKI2019)." *Procedia Computer Science* 174, 433–42.

United States Census Bureau. (2019). "LEHD Origin-Destination Employment Statistics (LODES) Dataset Structure Format Version 7.4." Retrieved from: https://lehd.ces.census.gov/data/lodes/LODES7/LODESTechDoc7.4.pdf.

Walsh, E. S., B. J. Kreakie, M. G. Cantwell, and D. Nacci. (2017). "A Random Forest Approach to Predict the Spatial Distribution of Sediment Pollution in an Estuarine System." *PLoS One* 12(7), e0179473. https://doi.org/10.1371/journal.pone.0179473.

Weinberger, R. R. (2001). "Light Rail Proximity: Benefit or Detriment in the Case of Santa Clara County, California?" *Transportation Research Record* 1747, 104–11.

Weinstein, B., and T. L. Clower. (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.

Yan, B., K. Janowicz, G. Mai, and R. Zhu. (2019). "A Spatially Explicit Reinforcement Learning Model for Geographic Knowledge Graph Summarization." *Transactions in GIS* 29(3), 620–40.