Big Data, Accessibility, and Urban House Prices

Steven C. Bourassa

Department of Urban and Regional Planning, Florida Atlantic University, 777 Glades Road, SO 284, Boca Raton, FL 33431, USA, sbourassa@fau.edu

Martin Hoesli

Geneva School of Economics and Management and Swiss Finance Institute, University of Geneva, 40 boulevard du Pont-d'Arve, CH-1211 Geneva 4, Switzerland, martin.hoesli@unige.ch, and School of Business, University of Aberdeen, UK

Louis Merlin

Department of Urban and Regional Planning, Florida Atlantic University, 777 Glades Road, SO 284, Boca Raton, FL 33431, USA, lmerlin@fau.edu

John Renne

Department of Urban and Regional Planning, Florida Atlantic University, 777 Glades Road, SO 284, Boca Raton, FL 33431, USA, jrenne@fau.edu

Big Data, Accessibility, and Urban House Prices

Abstract

Big data applications are attracting increasing interest on the part of urban researchers. One unexplored question is whether the inclusion of big data accessibility indexes improves the accuracy of hedonic price models used for residential property valuation. This paper compares a big data index with an index derived from a regional travel demand model developed by local transportation planning agencies and traditional measures of accessibility defined as distances to employment centers. Controls for submarkets and a combined spatial autoregressive and spatial error model are also assessed as tools for capturing the value of location. Using single-family residential transactions from the Miami, Florida, metropolitan area, the study's main conclusion is that the big data accessibility measure does not add meaningful explanatory or predictive power. In contrast, the spatial autoregressive and error model outperforms the other options considered.

Keywords

Big data, accessibility indexes, hedonic models, spatial models, property valuation

Introduction

The widespread use of mobile communication devices creates new data sources, particularly related to the movement of people, that are useful to urban researchers. Firms now provide access to this type of big data, which is based on the movements of millions of mobile phone users. Previously, the National Household Travel Survey (NHTS) was the only major national U.S. source of detailed travel behavior data; however, during the past two decades the NHTS has been conducted only in 2001, 2009, and 2017. The 2017 NHTS collected data on the self-reported movements of participants across 129,696 households over a one-day period (Westat, 2018). Travel surveys of metropolitan regions are conducted only every few years at most. For example, the most recent travel survey in Southeast Florida was conducted in 2017, but the most recent one before that was in 2000, meaning that comprehensive travel information was updated only once in the past 20 years. In comparison, big data from mobile communications can report the actual movements of millions of people on an ongoing basis. These data can be used for unprecedentedly detailed analysis of travel behavior. The NHTS and other travel diary surveys have some advantages for researchers, including detailed information collected about individuals and their households, as well as trip purposes. However, big data have the advantage of providing much greater spatial and temporal detail, larger sample sizes, and recency. These are the key defining characteristics of big data, which are commonly referred to as the three V's: variety, volume, and velocity (Laney, 2001; Winson-Geideman et al., 2018).

In recent years, a number of studies have applied indicators derived from big data sources to urban research, including topics such as smart cities (Al Nuaimi et al., 2015; Batty, 2013; Lim et al., 2018), planning practice and sustainability (Hao et al., 2015; Kitchin, 2014), and transportation (Cottrill and Derrible, 2015; Tao et al., 2014; Toole et al., 2015). Some studies have applied big data sources to real estate research (Glaeser et al., 2018; Winson-Geideman et al., 2018), for example to better inform local housing price dynamics on a national scale for all zip codes and census tracts over a 40-year period (Bogin et al., 2019). In practice, big data is used for a variety of real estate applications, such as in support of Zillow's well-known online property valuation service.

Hedonic house price studies typically use measures of distance or proximity to employment centers and transportation infrastructure to approximate accessibility to key destinations (Debrezion et al., 2007; Haider and Miller, 2000). Some of this research uses gravity models, network distances, or travel costs (Martínez and Viegas, 2009; Ryan, 1999). Other studies use dummy variables for geographic submarkets, which capture accessibility to employment as well as other aspects of location and neighborhood quality (Bourassa et al., 2003). Studies have also examined how property values respond as transport systems reshape accessibility in cities (Mulley, 2014; Shyr et al., 2013). Other research has focused on pedestrian accessibility indexes calculated for small areas from published indexes such as Walk Score (Boyle et al., 2014; Guo et al., 2017). Finally, much urban housing and real estate research has focused on spatial models that use information on neighboring properties as a way of capturing locational effects in a hedonic model (Bourassa et al., 2007; Can, 1992).

The focus of this study is to examine the benefits for urban residential property valuation of using an employment accessibility index derived from a big data source known as Sugar Access, obtained from the firm Citilabs. The Sugar Access data that underlie our index satisfy the criteria of variety, volume, and velocity, while traditional accessibility indexes lack the characteristics of velocity and volume as the data are collected only periodically and are based on a sample of a few thousand households at most. The present study compares a big data index with: an index derived from a regional travel demand model used by local transportation planning agencies; traditional measures of locational value based on straight-line distances to employment centers; controls for submarkets; and spatially lagged regressors and error terms. We compare hedonic model prediction accuracy results across the various specifications. The study uses sales data from 57,130 single-family home sales transactions across the three counties of the Miami, Florida, metropolitan area in 2016.

The various methods used to specify location in hedonic house price models reflect competing explanations in the literature on households' residential location choice. In the Alonso-Muth-Mills monocentric model, urban house prices are inversely related to the cost of commuting to employment at a central location (Alonso, 1964). Measures of distance to a central business district (CBD) capture this relationship. This model can easily be extended to allow for multiple centers of employment by including measures of distance to employment subcenters. At its extreme, the model results in the use of employment accessibility indexes to account for the location of all employment, whether in a center, subcenter, or elsewhere (Ahlfeldt, 2011). The Alonso-Muth-Mills model has been used to explain the pattern found in many cities where wealthier households tend to live further away from the center: in such cases, the desire to consume more land and housing outweighs the increase in commuting costs.

However, researchers such as Wheaton and Hamilton have questioned this model and its implications for residential location choice. Wheaton (1977, p. 631) finds 'that the suburbanization of America's middle and upper classes is a response to housing market externalities and the fiscal incentives of municipal fragmentation.' In other words, residential location choices of wealthier households have more to do with avoidance of disamenities, attraction to amenities, and preference for an optimal mix of local public services and taxes (Tiebout, 1956); other households are constrained to locate where wealthier households do not want to live. Hamilton (1982) and others (Giuliano and Small, 1993) focus on what Hamilton calls 'wasteful commuting', which refers to the fact that on average workers commute far more than would be predicted by models based on accessibility to employment. These ideas are reflected in hedonic house price models that include dummy variables for submarkets or information about nearby properties in the form of spatial lags. These variables of course also indirectly capture employment accessibility. To the extent that other factors play a larger role in residential location decisions in our study area, the models with submarket dummies or spatial lags will perform better than those with just employment accessibility measures.

This paper contributes to the literature on both hedonic modeling and big data by testing the usefulness of an employment accessibility index based on big data automobile travel times collected from personal mobile communication devices. Given the greater granularity (i.e., volume) and the improved recency (i.e., velocity) of the data, our hypothesis was that the big data index would yield hedonic models with greater predictive accuracy. However, this turns out to be a case where the contribution of big data is at best marginal as we find that the big data index contributes little or no explanatory or predictive power to the hedonic model. In contrast to the big data accessibility index, a combined spatial autoregressive and spatial error model performs best, consistent with the importance given to such models in the urban housing and real estate literature. This result is also consistent with the Wheaton-Hamilton critique of the Alonso-Muth-Mills model.

The next section presents our empirical strategy. The subsequent two sections discuss our data and results, respectively. A final section provides some concluding remarks.

Empirical strategy

Study area

Our study area consists of Miami-Dade, Broward, and Palm Beach counties in southeastern Florida. The three counties together form the Miami-Fort Lauderdale-West Palm Beach, FL Metropolitan Statistical Area (the Miami MSA), which is currently ranked seventh in the United States with an estimated 2019 population of nearly 6.2 million (US Census Bureau, 2019a). The urbanized part of the MSA is situated between the Atlantic Ocean to the east and the Everglades to the west and is narrow from east to west (approximately 5 to 20 miles), but lengthy from north to south (approximately 100 miles). The MSA has multiple limited access expressways that run either north and south or east and west. It is also served by three county-specific public transit agencies that provide bus services and, in Miami-Dade County, a commuter rail system as well as an automated train system that operates in the Miami downtown. There is also a publicly operated commuter rail system that connects the three counties but does not directly serve key downtown areas.[1] There are three commercial airports, located in Miami, Fort Lauderdale, and West Palm Beach.

The Miami MSA might seem an odd choice for a study of employment accessibility indexes due to the relatively large numbers of retirees and second homes, implying that the number of workers per household might be low. In fact, the Miami MSA has a high average number of workers per household,

[1] A new private passenger rail service (which began in 2018 but was halted in March 2020 due to COVID-19) connects the downtowns of Miami, Fort Lauderdale, and West Palm Beach, with plans to extend north to Orlando.

ranking at the 88th percentile of the distribution of the number of workers per household across all MSAs according to the American Community Survey 2012-2017 estimates (US Census Bureau, 2020); this is similar to the New York or San Francisco MSAs.

Hedonic models

We use hedonic models to relate transaction prices for single-family houses to the characteristics of those houses, including a range of measures of aspects of the structure and lot as well as measures of location, including accessibility indexes. Our model specifications are constrained by the variables available in the county assessors' datasets as well as our ability to calculate variables related to amenities, disamenities, and employment accessibility or proximity. We start with a simple hedonic model that has only structural and lot characteristics, amenities and disamenities, school quality ratings, and monthly time dummies; then we add different types of variables that relate to employment accessibility.

A primary focus for the study is the analysis of how access to employment influences the prices of single-family homes. We compare three approaches to measuring this directly: two accessibility indices and distances to employment centers. One employment accessibility index is calculated using the regional travel demand model used by local transportation planning agencies, which is based on surveys of travel behavior; although this model was built in 2017, it is calibrated to travel behavior from the year 2009 (WSP Parsons Brinckerhoff, 2017). The other index is calculated using Citilabs' Sugar Access data, which is based on travel time information collected from personal devices, such as cell phones.[2] As discussed in the Introduction, the Sugar Access index fits the definition of big data better than the travel demand model index with respect to the velocity and volume criteria, given that the travel time and employment destination location are updated annually (and could in theory be updated in real time) and are based on millions of trips. In contrast, the survey data used for the travel demand model are collected only infrequently and are based on much smaller samples of trips; 19,630 trips were collected in the 2017 survey (WSP Parsons Brinckerhoff, 2017). These two indices are compared with distances to employment centers.[3] We measure the straight-line distance to the nearest CBD or employment

[2] We focus on automobile accessibility given that most commutes in the Miami MSA are by personal automobile; only 3.5% of commutes are by public transportation, according to the 2014-2018 American Community Survey (US Census Bureau, 2020).

[3] We also considered distances to transportation nodes, such as rail stations and expressway interchanges, but these were collinear with distance to the CBD or disamenity measures and so were not included in the models reported here.

subcenter. The latter can be the nearest CBD, the nearest major airport, or another area with a significant concentration of employment.[4]

We also examine two other modeling approaches that account for spatial differences without measuring access to employment directly. One approach is to create dummy variables for submarkets, and the other is to employ spatial modeling techniques. Previous research shows that submarket dummies can add significantly to the explanatory power of hedonic models by capturing accessibility as well as neighborhood effects that are otherwise not measurable (Bourassa et al., 2003). Here we produce a set of area dummy variables by combining nearby census tracts with the aim of having at least 100 observations per area. Combining tracts avoids the potential problem of overfitting due to small numbers of observations. This resulted in 178 areas across the three counties in the Miami MSA.

Finally, we estimate a model with spatial autoregressive and spatial error terms as an alternative means for capturing locational effects. Spatial dependence in urban housing and real estate markets has been a topic of study for several decades (Basu and Thibodeau, 1998; Can, 1992; Pace et al., 1998). Following Anselin (2003), we initially specified a model with a spatially lagged dependent variable and a spatially lagged error term. However, maximum likelihood estimators typically used for spatial modeling did not converge for our samples, possibly due to the large number of observations. Hence, we followed Anselin's suggestion and estimated a model using OLS with lagged control variables as regressors along with the original independent variables. This approach avoids the endogeneity that would result from including the spatially lagged dependent variable in an OLS model. We also include the lagged error term in the OLS estimation because that does not lead to biased estimates, although it does result in biased standard errors (Anselin, 2003). Consequently, the significance levels we report for the variables in this model may in some cases be incorrect. Given our primary focus on prediction accuracy, which is not dependent on precise standard errors, this is not an issue.

Model comparison

We estimate multiple models to assess how the different variables that are related to employment accessibility or proximity contribute to explanatory power and prediction accuracy. All models contain the lot size and structural variables, amenities and disamenities, public school ratings, and time

[4] Including airports as employment subcenters allows us to disentangle the positive and negative impacts of airports (in the model that includes the employment subcenter variable). The other employment subcenters are Coral Gables, Kendall, and Miami Beach (in Miami-Dade County), Sunrise and Cypress Creek (in Broward County), and Boca Raton (in Palm Beach County).

dummies. We first estimate a model with just those variables. Then we estimate models that add the travel demand model automobile accessibility index, the Sugar Access index, distances to employment centers, or area dummies. Finally, a model that contains spatially lagged control variables and residuals is assessed as a tool for capturing the value of location. This results in a set of six estimations.

We compare adjusted *R*-squared statistics, which measure in-sample explanatory power, as well as several measures of out-of-sample prediction accuracy, which is more directly pertinent to property valuation. We assign the transactions data to 10 random samples each containing 80% of the data. We use these samples for estimation purposes. Then we predict house prices and calculate the following accuracy statistics for the remaining 20% samples: mean absolute error, mean percentage error, and the proportions of predictions within 10% or 20% of the actual prices. This approach ensures that the results do not depend on the sample used for estimation purposes.

Data

Transactions data and control variables

We use single-family house transactions data for 2016 collected by property tax assessors for Miami-Dade, Broward, and Palm Beach counties. The data were obtained from the Florida Department of Revenue, which maintains property valuation data for all Florida counties as part of their property tax oversight efforts. We exclude extremely high- and low-priced as well as large and small properties (the top and bottom 1% in each case), resulting in a sample of 57,130 transactions across the three counties.

We use the following variables from the transactions data: sale price, size of the lot, floor area, year of construction, building quality, the value of special features, and the month of sale. The dependent variable is the natural logarithm of sale price, which is a common specification in hedonic models due to the need to normalize the skewed distribution of house prices. The lot size, floor area, and value of special features are specified in logarithmic terms to reflect diminishing returns. We calculate property age by subtracting the construction year from the year of the transaction. Age enters the model in quadratic form to reflect the initial decline in value of the structure, followed by renovation and an increase in value. Building quality is rated on a six-point scale that we convert to four dummy variables because not all categories are used in each county and in some cases have few observations. We include monthly dummy variables to capture price trends. Published indexes, such as those produced by the Federal Housing Finance Agency, show price increases in the Miami region during 2016 (FHFA, 2020).

We supplement the transactions data with measures of various amenities and disamenities as well as school quality measures. The amenities are proximity to the ocean and the nearest body of water, while the disamenities are airport noise and proximity to the nearest railroad and expressway. The distance variables are calculated as straight-line distances using GIS software. The airport noise data come from the Bureau of Transportation Statistics (2020) and reflect equivalent sound levels (LEQs). These levels reflect averages over 24 hours for a day in 2018 and are slightly lower than day-night average sound levels (DNLs), which add a penalty to nighttime noise. According to the Federal Aviation Administration (2020), DNLs in excess of 65 are unsuitable for residential neighborhoods. To adjust for the nighttime penalty, we classify LEQs in excess of 60 as unsuitable for residential areas and specify a dummy variable accordingly (Cohen and Coughlin, 2008).[5] Distances to amenities and disamenities are measured in logarithmic terms to reflect diminishing effects.

School quality measures have been shown to have significant effects on house values (Downes and Zabel, 2002; Haurin and Brasington, 1996) and hence we include dummy variables for elementary, middle, and high school grades as controls. Ries and Somerville (2010) conclude that some of the findings of previous studies may result from school quality measures picking up the effects of other unmeasured aspects of neighborhood quality, which could be the case in the present study. It is possible to control for the quality of public schools because they generally draw students from defined zones (although some of the better and magnet schools draw from outside their zones); this is not the case for private schools.[6] The school quality measures are from the Florida Department of Education (2019). Schools are evaluated each year on multiple criteria, such as student performance on standardized tests, and are graded A, B, C, D, or F based on the average of the points received for each of those criteria. Not all grades are reflected in each county.

Accessibility and employment proximity measures

We calculate accessibility to jobs with two different data sets, one from a regional travel demand model and the other from the big data provided via Citilabs' Sugar Access tool. The Southeast Florida Regional Planning Model (SERPM) simulates the full spectrum of a person's travel behavior over the course of a day (Parsons Brinckerhoff and The Corradino Group, 2016). The model builds a synthetic population

[5] We considered a similar variable for road noise but found that the variable measuring distance from the nearest expressway was better able to capture the effect of this disamenity.

[6] The quality of public schools in south Florida is generally good (Background Checks.org, 2019), suggesting that they are real alternatives to private schools and that public-school quality should have a measurable impact on property values.

with detailed demographic characteristics and provides outputs including trip origin, trip destination, trip time of day, and travel mode. These data are then synthesized across the population to derive congestion-adjusted travel times from zone to zone across the region.

The Sugar Access tool is proprietary software that builds on the capabilities of GIS to calculate accessibility measures. This tool can consider accessibility to destinations for automobiles, transit, biking, and walking, and can calculate accessibility to many kinds of destinations, namely employment and population centers and points of interest, which can be further disaggregated by type. It employs a gravity-type accessibility formula that assumes more distant destinations are less attractive than proximate destinations. The user can customize the zonal structure used to define origins and destinations within the software, though the total number of zones is limited to 32,000.

In each case, the calculation of accessibility requires a definition of origin zones, *i*, destination zones, *j*, a measure of attractiveness for each destination, d_j, and a measure of travel cost from *i* to *j*, c_{ij}. We adopt a gravity (or potential) accessibility measure for each zone, A_i:

$$A_i = \sum_j d_j e^{-\beta c_{ij}} \tag{1}$$

A large body of research makes use of gravity-based accessibility measures because they balance a strong theoretical and behavioral basis with a simple technical expression (Grengs et al., 2010; Hu, 2017).[7] We employ the negative exponential formulation of the gravity formula, which has been demonstrated to accurately reflect commuting travel behavior in a wide variety of metropolitan settings The formula essentially gives more distant destinations less weight, because such destinations are less attractive to a worker.

To calculate A_i, destination attractiveness d_j is measured in total jobs and travel cost c_{ij} is measured in minutes. We use a spatial interaction model calibrated to the South Florida Household Travel Survey from 2017 and find that $\beta = 0.05911$ (units in inverse minutes) (WSP Parsons Brinckerhoff, 2017). In the case of the travel demand model, we account for the effects of tolling on route choice.[8] We convert tolls into minutes at 60% of the average wage in the region, so a $1 toll adds the equivalent of 6.6 minutes to travel time. We note that this may seem counter-intuitive because people often pay tolls for express

[7] Gravity-type job accessibility indicators more accurately describe commuting patterns and residential location choice than simpler alternatives such as cumulative opportunity measures (Bunel and Tovar, 2014; Geurs and van Wee, 2004).

[8] The Sugar Access tool is incapable of incorporating tolls; however, this is unlikely to affect our conclusions.

lanes to save time, but to calculate accessibility, either money must be converted to time or time to money and this study converts money into a time penalty.

Different input data are used for the travel demand and Sugar Access measures. The travel demand model provides 4,236 Travel Analysis Zones (TAZs).[9] For the Sugar Access measure, we define 29,217 zones based upon 28,626 census blocks covering the most densely developed parts of the region and 591 block groups for less developed areas. Total job counts for destinations come from the Longitudinal Employer-Household Dynamics (LEHD) Origin-Destination Employment Statistics for 2015 for the travel demand model and 2012 for Sugar Access (US Census Bureau, 2016).[10]

Zone-to-zone travel times for the travel demand model are predetermined by the model, whereas they are computed within the Sugar Access model using the most recent link-level speeds provided by the company HERE (which provides data for a total of 21 million roadway links).[11] The link-level speed data are available for four times of day: morning peak, afternoon, evening peak, and nighttime (Citilabs, 2018). We selected morning peak travel times for both the travel demand and Sugar Access measures. A comparison of the inputs used for the two accessibility calculations is in Table 1.

Figures 1a and 1b display automobile accessibility to jobs for the Miami MSA with equal interval maps, for the travel demand model and Sugar Access, respectively. In both cases, the highest job accessibility zones are in the City of Miami, with a zone of high accessibility extending northwards along the I-95 expressway corridor and including the City of Fort Lauderdale. West Palm Beach, the northernmost major city for the region, displays a more modest level of accessibility by automobile. Despite the differences between the two sources of data, the house-level correlation between the travel demand and Sugar Access automobile accessibility indexes is $\rho = 0.957$. Both indexes are expressed in logarithmic terms reflecting diminishing returns to accessibility.

As discussed above, the usefulness of the accessibility indexes is gauged against more traditional measures, namely distances to the nearest CBD and employment subcenter. Both distances are measured in logarithmic terms, reflecting the well-documented negative exponential relationship

[9] Rural areas of Miami-Dade and Broward counties are not covered by these zones.

[10] The LEHD data are aggregated by the US Census Bureau from state unemployment records, the Quarterly Census of Employment and Wages, Business Dynamics Statistics, and other demographic sources to derive the number of jobs, their location, their industry, and worker demographic characteristics such as age and gender. The data set covers 95% of private US employment and most federal jobs as well. Job locations are provided at the census block level (US Census Bureau, 2019b).

[11] HERE is a digital mapping and location data company that provides travel time data worldwide (see www.here.com).

between distance to employment centers and property values. We also assess the performance of the indexes against a set of dummy variables for submarkets, which are expected to capture proximity to employment as well as the impacts of other neighborhood characteristics. There are 68, 66, and 44 areas in Miami-Dade, Broward, and Palm Beach counties, respectively. Finally, we specify the spatial autoregressive and error model to include lags of all control variables except for the monthly dummy variables and lags of the residuals. The lags are based on the five nearest neighbors, each weighted by the inverse of the distance to the property in question.

Results

Preliminary analysis indicated that better results could be obtained by modeling each of the three counties separately. This is because certain variables (such as building and school quality) appear to be measured differently across counties and the impacts of other variables (such as distance to the CBD) vary across counties. Nevertheless, the employment accessibility measures and distances to the nearest CBD and employment subcenter are calculated without regard to county boundaries. Table 2 provides sample statistics for the largest county, Miami-Dade. The dummy variables for submarkets and the lagged independent variables for the spatial model are not included in the table.

Table 3 contains the specifications and prediction accuracy results for the six models for each county, while Table 4 gives estimation results for Miami-Dade County. Model 1 is the base model and contains only the control variables: lot and structural characteristics, distances to amenities and disamenities, school ratings, and monthly dummy variables. Models 2 through 6 each contain the control variables and one type of measure related to job accessibility: the automobile accessibility index based on the travel demand model in Model 2; the index based on Sugar Access data in Model 3; the two distances to employment centers in Model 4; the area dummy variables in Model 5; and the spatial lags in Model 6.

When only the control variables are included in the Miami-Dade County model, the adjusted *R*-squared is 0.809 (Model 1). A small increase in the adjusted *R*-squared (to 0.817) is obtained when the travel demand accessibility index is added (Model 2). A similar result (0.819) is achieved when the Sugar Access index is included (Model 3). Distances to employment centers (Model 4) are more effective than either of the automobile accessibility indexes (0.840). The submarket dummies (Model 5) perform notably better than the accessibility indexes or distances to employment centers (0.897). Finally, the spatial lags

(Model 6) clearly work the best (0.935).[12] For Broward and Palm Beach counties, where jobs are less centralized compared to Miami-Dade county, the accessibility indexes do not add any explanatory power relative to the base model. Similarly, the distances to employment centers add little or nothing to the models for those two counties. As is the case for Miami-Dade County, the addition of submarket dummies increases the adjusted *R*-squared (from 0.825 for the base model to 0.872 for Broward County and from 0.813 to 0.847 for Palm Beach County), while the best results are achieved with the combined spatial autoregressive and spatial error model (with *R*-squared values of 0.909 and 0.921 for the two counties, respectively).

Turning to the prediction accuracy results in Table 3, the usefulness of the spatial autoregressive and spatial error approach appears again, while the accessibility indexes contribute only marginally. For instance, in Miami-Dade County 87.4% of predictions are within 20% of the transaction price when the spatially lagged variables are included in the model, compared to only 63.5% for the base model. The comparable figures for Models 2 through 5 are 66.0%, 66.4%, 70.3%, and 80.5%, respectively.

As shown in Table 4 with respect to Miami-Dade County, the control variable coefficients are in most cases significant with expected signs. The coefficients for land and floor area are both positive and highly significant (at better than the 1% level) across all models. Age and age squared are negative and positive, respectively, consistent with the cycle of depreciation and reinvestment. The estimates for the average construction quality variable are unexpectedly higher than those for the two better quality categories suggesting that this variable may be picking up some aspect of location value. The value of special features is positively related to property value. With respect to amenities and disamenities, the distances to the ocean and the nearest water body have the expected negative and significant coefficients. Distances to the nearest railroad and expressway are, with one exception, positively related to value, reflecting the noise or perceived negative air quality produced by those features. The coefficient on airport noise is negative and significant in Models 4 and 6, presumably because Model 4 also measures proximity to the nearest employment subcenter, which could be an airport, and Model 6 does a better job of controlling for locational effects than the other models. The school quality indicators do not always show higher values for better quality, implying that the school zones are measuring neighborhood characteristics as well as school quality, consistent with the findings of Ries and Somerville (2010). Another factor could be the move towards an open enrollment public school

[12] Combining different types of accessibility measures in the same model has virtually no impact on the adjusted *R*-squared statistics but often causes multicollinearity problems. This is also true when the spatial autoregressive and spatial error terms are included in the model along with other accessibility measures.

system where location and school are less connected, similar to private schools. The monthly dummy variables generally show a pattern of increasing values as the year progresses, consistent with other price indexes such as those produced by FHFA.

The automobile accessibility indexes are in each case highly significant with the anticipated positive sign. The distances to employment centers have the expected negative sign and are also highly significant. Consistent with the statistics in Table 3, the relationship between employment accessibility or proximity and property values is less pronounced in Broward and Palm Beach counties. As noted above, these counties are more suburbanized than Miami-Dade County. The Census's OnTheMap tool, which shows the geographic distribution of jobs, reveals that Downtown Miami and Miami Beach accounted for 25% of all jobs in Miami-Dade County in 2017. Downtown Fort Lauderdale and Fort Lauderdale Beach accounted for 10% of the jobs in Broward County and Downtown West Palm Beach and Palm Beach accounted for only 7% of the total jobs in Palm Beach County (in all three counties, job clusters are present in the CBD and nearby beachfront, which is located within a 10-minute drive of the CBD).

For Broward County, the estimate for the travel demand index in Model 2 is positive but significant at only the 10% level, while the estimate for the Sugar Access index in Model 3 is not significant. In Model 4, the distance to the CBD is significant at the 1% level, but the distance to the nearest employment subcenter is not significant. For Palm Beach County, the estimate for the travel demand index is also positive and significant at the 10% level, while the estimate for the Sugar Access index is positive and significant at the 1% level. For Model 4, the distance to the CBD is significant at the 1% level but with the wrong sign, while the distance to the nearest employment subcenter is positive and significant at the 1% level. The latter result appears to reflect the fact that employment accessibility in Palm Beach County is greatest in the southern part of the county (around the Boca Raton employment subcenter) rather than near West Palm Beach.

Conclusions

For the Miami metropolitan area, we find that access to employment has a marginal influence at best in predicting single-family home prices, especially in heavily suburbanized counties where jobs are decentralized. This is true for all three employment accessibility measures used, including our big-data job accessibility measure. Models with dummy variables for geographical submarkets are more useful than any of the employment accessibility models in all three counties, evidently at least in part because they capture other valued neighborhood characteristics as well as proximity to employment. A spatial model that includes both autoregressive and error terms outperforms the others. These results are consistent with the Wheaton-Hamilton critique of the Alonso-Muth-Mills model.

The results of this study are testimony to the fact that, in the context of hedonic modeling, more data and more granular data are not necessarily better data. While our paper provides an example of the ineffectiveness of big data for employment accessibility in hedonic models, the potential of such data to better understand and predict property values should not be ruled out entirely. Given the study area, especially Broward and Palm Beach counties where the role of the downtown is weak, we cannot generalize this finding to other metropolitan areas that are more centralized. The finding in Miami-Dade County provides some evidence that employment accessibility measures may in fact have some explanatory power in counties with more centralized employment. More research is needed across more regions to further examine this topic.

Another direction for future study is to apply big data to more detailed analysis of movement (i.e., not just employment-related travel) to understand what aspects of proximity or distance matter with respect to house values. Can we observe patterns of movement in high-value neighborhoods that explain their greater value? Are certain amenities or disamenities hidden within the structure of metropolitan areas that could be uncovered by a deeper understanding of how residents circulate? Future research could also examine the relationship of employment accessibility and property value by type of job and industry cluster. For example, Boca Raton is noted for having a cluster of high-tech jobs and higher property values. On the other hand, service industry jobs, which dominate South Florida due to the tourism industry, may have less impact on real estate prices.

In a nutshell, our results demonstrate that, in a relatively decentralized metropolitan area, accessibility measures based on big data add little or no predictive power to hedonic models. Of course, our conclusions say nothing about the value of big data on travel behavior for purposes other than hedonic price prediction. The kind of data that underlie our accessibility indexes is used every day by billions of people around the world to choose travel routes, estimate travel times, and avoid delays, all of which clearly enhances the wellbeing of urban travelers.

Acknowledgements

The authors thank Andrew VanValin, Estefania Mayorga, and Jessica Kluttz for helpful research assistance. We also thank the reviewers and editors for helping us to improve the paper.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Land Economics Foundation.

References

Ahlfeldt, G (2011) If Alonso was right: Modeling accessibility and explaining the residential land gradient. *Journal of Regional Science* 51(2): 318-338.

Al Nuaimi E, Al Neyadi H, Mohamed N and Al-Jaroodi J (2015) Applications of big data to smart cities. *Journal of Internet Services and Applications* 6(25): 1-15.

Alonso W (1964) *Location and Land Use*. Cambridge, MA: Harvard University Press.

Anselin L (2003) Spatial econometrics. In: Baltagi BH (ed) *A Companion to Theoretical Econometrics*. Malden, MA: Blackwell, pp. 310-330.

Background Checks.org (2019) Top 100 school districts in America 2019. Available at: backgroundchecks.org/top-school-districts.html (accessed 27 March 2020).

Basu S and Thibodeau TG (1998) Analysis of spatial autocorrelation in housing prices. *Journal of Real Estate Finance and Economics* 17(1): 61-85.

Batty M (2013) Big data, smart cities and city planning. *Dialogues in Human Geography* 3(3): 274-279.

Bogin A, Doerner W and Larson W (2019) Local house price dynamics: New indices and stylized facts. *Real Estate Economics* 47(2): 365-398.

Bourassa SC, Cantoni E and Hoesli M (2007) Spatial dependence, housing submarkets, and house price prediction. *Journal of Real Estate Finance and Economics* 35(2): 143-160.

Bourassa SC, Hoesli M and Peng VS (2003) Do housing submarkets really matter? *Journal of Housing Economics* 12(1): 12-28.

Boyle A, Barrilleaux C and Scheller D (2014) Does walkability influence housing prices? *Social Science Quarterly* 95(3): 852-867.

Bunel M and Tovar E (2014) Key issues in local job accessibility measurement: Different models mean different results. *Urban Studies* 51(6): 1322-1338.

Bureau of Transportation Statistics (2020) *CONUS Aviation Noise Image Service*. Available at: osav-usdot.opendata.arcgis.com/datasets/ (accessed 5 June 2020).

Can A (1992) Specification and estimation of hedonic housing price models. *Regional Science and Urban Economics* 22(3): 453-474.

Citilabs (2018) *Sugar Access User Guide*. Sacramento, CA.

Cohen JP and Coughlin CC (2008) Spatial hedonic models of airport noise, proximity, and housing prices. *Journal of Regional Science* 48(5): 859-878.

Cottrill CD and Derrible S (2015) Leveraging big data for the development of transport sustainability indicators. *Journal of Urban Technology* 22(1): 45-64.

Debrezion H, Pels E and Rietveld P (2007) The impact of railway stations on residential and commercial property value: A meta-analysis. *Journal of Real Estate Finance and Economics* 35(2): 161-180.

Downes TA and Zabel JE (2002) The impact of school characteristics on house prices: Chicago 1987-1991. *Journal of Urban Economics* 52(1): 1-25.

Federal Aviation Administration (2020) *Fundamentals of Noise and Sound*. Available at: www.faa.gov/regulations_policies/policy_guidance/noise/basics/ (accessed 5 June 2020).

Federal Housing Finance Agency (2020) *FHFA House Price Index*. Available at: www.fhfa.gov/DataTools/Downloads/Pages/House-Price-Index.aspx (accessed 30 March 2020).

Florida Department of Education (2019) *Florida School Accountability Reports*. Available at: www.fldoe.org/accountability/accountability-reporting/school-grades/ (accessed 3 June 2019).

Geurs KT and van Wee B (2004) Accessibility evaluation of land-use and transport strategies: Review and research directions. *Journal of Transport Geography* 12(2): 127-140.

Giuliano G and Small KA (1993) Is the journey to work explained by urban spatial structure? *Urban Studies* 30(9): 1485-1500.

Glaeser EL, Kominers SD, Luca M et al. (2018) Big data and big cities: The promises and limitations of improved measures of urban life. *Economic Inquiry* 56(1): 114-137.

Grengs J, Levine J, Shen Q et al. (2010) Intermetropolitan comparison of transportation accessibility: Sorting out mobility and proximity in San Francisco and Washington, DC. *Journal of Planning Education* *and Research* 29(4): 427-443.

Guo Y, Peeta S and Somenahalli S (2017) The impact of walkable environment on single-family residential property values. *Journal of Transport and Land Use* 10(1): 241-261.

Haider M and Miller EJ (2000) Effects of transportation infrastructure and location on residential real estate values: Application of spatial autoregressive techniques. *Transportation Research Record* 1722: 1-8.

Hamilton BW (1982) Wasteful commuting. *Journal of Political Economy* 90(5): 1035-1053.

Hao J, Zhu J and Zhong R (2015) The rise of big data on urban studies and planning practices in China: Review and open research issues. *Journal of Urban Management* 4(2): 92-124.

Haurin D and Brasington D (1996) School quality and real house prices: Inter- and intrametropolitan effects, *Journal of Housing Economics* 5(4): 351-368.

Hu L (2017) Job accessibility and employment outcomes: Which income groups benefit the most? *Transportation* 44(6): 1421-1443.

Kitchin R (2014) The real-time city? Big data and smart urbanism. *GeoJournal* 79(1): 1-14.

Laney, D (2001) 3-D data management: Controlling data volume, velocity, and variety. *Application Delivery Strategies*. Stamford, CT: Meta Group.

Lim C, Kim KJ and Maglio PP (2018) Smart cities with big data: Reference models, challenges, and considerations. *Cities* 82: 86-99.

Martínez LM and Viegas JM (2009) Effects of transportation accessibility on residential property values: Hedonic price model in the Lisbon, Portugal, metropolitan area. *Transportation Research Record* 2115: 127-137.

Mulley C (2014) Accessibility and residential land value uplift: Identifying spatial variations in the accessibility Impacts of a bus transitway. *Urban Studies* 51(8): 1707-1724.

Pace RK, Barry R and Sirmans CF (1998) Spatial statistics and real estate. *Journal of Real Estate Finance and Economics* 17(1): 5-13.

Parsons Brinckerhoff and The Corradino Group (2016) *Southeast Florida Regional Planning Model, SERPM 7.0: Model Users Guide*.

Ries J and Somerville T (2010) School quality and residential property values: Evidence from Vancouver rezoning. *Review of Economics and Statistics* 92(4): 928-944.

Ryan S (1999) Property values and transportation facilities: Finding the transportation-land use connection. *Journal of Planning Literature* 13(4): 412-427.

Shyr O, Andersson DE, Wang J et al. (2013) Where do home buyers pay most for relative transit accessibility? Hong Kong, Taipei and Kaohsiung compared. *Urban Studies* 50(12): 2553-2568.

Tao S, Corcoran J, Mateo-Babiano I et al. (2014) Exploring Bus Rapid Transit passenger travel behaviour using big data. *Applied Geography* 53: 90-104.

Tiebout CM (1956) A pure theory of local expenditures. *Journal of Political Economy* 64(5): 416-424.

Toole JL, Colak S, Sturt B et al. (2015) The path most traveled: Travel demand estimation using big data resources. *Transportation Research Part C: Emerging Technologies* 58(Part B): 162-177.

US Census Bureau (2016) LEHD Origin-Destination Employment Statistics. Available at: lehd.ces.census.gov/data/lodes/LODES7/ (accessed 2 December 2016).

US Census Bureau (2019a) Population and Housing Unit Estimates. Available at: www.census.gov/programs-surveys/popest/data/tables.html (accessed 31 March 2020).

US Census Bureau (2019b) Quarterly Workforce Indicators 101. Available at: lehd.ces.census.gov/doc/QWI_101.pdf (accessed 31 March 2020).

US Census Bureau (2020) American Community Survey. Available at: www.census.gov/data.html (accessed 19 June 2020).

Westat (2018) 2017 NHTS Data User Guide. Submitted to the Federal Highway Administration. Available at: nhts.ornl.gov/assets/2017UsersGuide.pdf (accessed 15 May 2019).

Wheaton WC (1977) Income and urban residence: An analysis of consumer demand for location. *American Economic Review* 67(4): 620-631.

Winson-Geideman K, Krause A, Lipscomb, CA et al. (2018) *Real Estate Analysis in the Information Age: Techniques for Big Data and Statistical Modeling*. New York: Routledge.

WSP Parsons Brinckerhoff (2017) *Southeast Florida Household Travel Survey*. Miami.

Table 1. Characteristics of accessibility calculations from travel demand and Sugar Access tools.

Model features	Travel demand	Sugar Access
Number of zones (three counties combined)	4,236	29,217
Job data	LODES 2015 All Jobs	LODES 2012 All Jobs
Travel time	SERPM (2010 morning peak, toll costs included)	HERE Technologies (2016 morning peak, toll costs not included)
Impedance	Negative exponential (calibrated to 2017 commutes)	Negative exponential (calibrated to 2017 commutes)

Notes: LODES = LEHD Origin-Destination Employment Statistics; SERPM = Southeast Florida Regional Planning Model.

Table 2. Descriptive statistics for Miami-Dade County sample.

Variable	Mean or percentage	Standard deviation	Minimum	Maximum	Transformation or default category
Sale price ($)	399,942	317,215	72,000	2,650,000	ln
Land area (sq. ft.)	8,621	6,070	1,248	57,064	ln
Structural characteristics					
Floor area (sq. ft.)	2,058	814	854	6,287	ln
Age	30.7	21.2	0.0	96.0	quadratic
Quality (%)					
Minimum or below average	30.8	—	0	1	default category
Average	0.1	—	0	1	
Above average	54.7	—	0	1	
Excellent or superior	14.4	—	0	1	
Special features' value ($)	9,562	13,891	0	175,020	ln
Amenities and disamenities					
Distance to ocean (ft.)	31,691	17,595	236	75,745	ln
Distance to nearest water body (ft.)	11,960	11,933	0	50,400	ln
Distance to nearest railroad (ft.)	8,349	6,178	10	29,622	ln
Distance to nearest expressway (ft.)	7,724	6,069	90	48,167	ln
Aviation noise > 60 LEQ (%)	1.5	—	0	1	
School quality (%)					
Elementary school					
A	41.9	—	0	1	
B	27.1	—	0	1	
C	27.3	—	0	1	
D	3.7	—	0	1	default category
Middle school					
A	13.3	—	0	1	
B	25.0	—	0	1	
C	45.8	—	0	1	
D	15.9	—	0	1	default category
High school					
B	25.7	—	0	1	
C	71.3	—	0	1	
D	3.0	—	0	1	default category
Month of sale (%)					
January	6.0	—	0	1	default category
February	6.7	—	0	1	
March	8.8	—	0	1	
April	8.9	—	0	1	
May	8.9	—	0	1	
June	10.0	—	0	1	
July	8.7	—	0	1	
August	9.2	—	0	1	
September	8.7	—	0	1	
October	7.6	—	0	1	
November	8.3	—	0	1	
December	8.3	—	0	1	

Variable	Mean or percentage	Standard deviation	Minimum	Maximum	Transformation or default category
Accessibility indexes					
Travel demand automobile	264,573	142,381	0	550,897	ln
Sugar Access automobile	276,071	112,818	48,690	502,889	ln
Distances to employment centers (ft.)					
Nearest CBD	68,490	32,008	3,826	159,977	ln
Nearest employment center	41,115	22,162	1,463	110,554	ln

Notes: *n* = 13,932. Means for dummy variables are converted to percentages. This table excludes statistics for the submarket dummy variables, school grade categories for which there are no observations (elementary school grade F and high school grade A), and spatially lagged variables.

Table 3. Model specifications and prediction accuracy results.

			Model			
	1	2	3	4	5	6
Model specification						
Control variables	√	√	√	√	√	√
Auto accessibility index based on travel demand model		√				
Auto accessibility index based on Sugar Access data			√			
Distances to employment centers				√		
Submarket dummies					√	
Spatially lagged control variables and error term						√
Miami-Dade County						
Adjusted R^2 (full sample)	0.809	0.817	0.819	0.840	0.897	0.935
Out-of-sample prediction accuracy						
Mean absolute error	79,519	76,123	75,634	69,771	55,923	42,760
Mean percentage error	19.5	18.9	18.7	17.2	13.3	10.3
Percent within 10%	33.8	36.3	36.7	39.9	52.1	63.8
Percent within 20%	63.5	66.0	66.4	70.3	80.5	87.4
Broward County						
Adjusted R^2 (full sample)	0.825	0.825	0.825	0.827	0.872	0.909
Out-of-sample prediction accuracy						
Mean absolute error	55,502	55,528	55,480	55,131	46,946	37,728
Mean percentage error	17.0	17.0	17.0	17.0	14.3	11.7
Percent within 10%	40.4	40.3	40.4	40.3	48.4	59.1
Percent within 20%	69.4	69.4	69.5	69.7	77.2	83.9
Palm Beach County						
Adjusted R^2 (full sample)	0.813	0.813	0.813	0.813	0.847	0.921
Out-of-sample prediction accuracy						
Mean absolute error	70,852	70,751	70,697	70,631	65,988	42,154
Mean percentage error	19.0	19.0	19.0	19.0	17.3	11.6
Percent within 10%	39.4	39.7	39.7	39.9	41.8	60.8
Percent within 20%	67.4	67.4	67.4	67.5	71.3	85.4

Note: √ indicates that the relevant variables are included in the model.

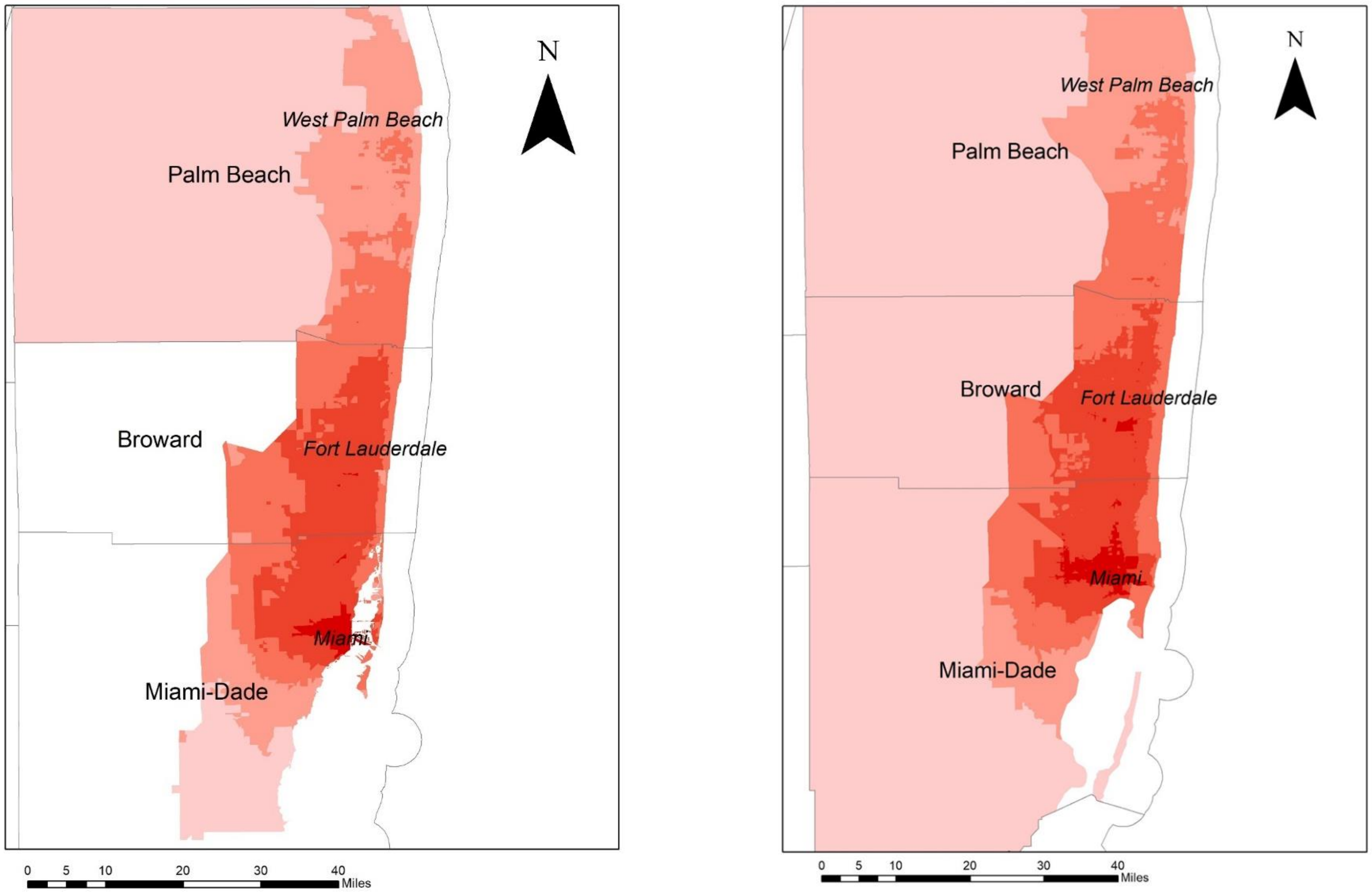
Table 4. Regression results for Miami-Dade County.

Variable	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Intercept	7.725***	5.805***	5.229***	9.365***	8.086***	7.602***
Land area (ln)	0.175***	0.184***	0.178***	0.161***	0.222***	0.241***
Structural characteristics						
Floor area (ln)	0.657***	0.680***	0.682***	0.686***	0.589***	0.454***
Age	-0.009***	-0.010***	-0.010***	-0.009***	-0.011***	-0.010***
Age squared	7.03x10^{-5}***	6.69x10^{-5}***	6.94x10^{-5***}	4.70x10^{-5***}	6.29x10^{-5***}	5.85x10^{-5***}
Quality						
Average	0.789***	0.688***	0.688***	0.671***	0.559***	0.452***
Above average	0.224***	0.198***	0.193***	0.178***	0.165***	0.120***
Excellent or superior	0.499***	0.438***	0.440***	0.404***	0.345***	0.273***
Special features' value (ln)	0.007***	0.006***	0.006***	0.006***	0.006***	0.005***
Amenities and disamenities						
Distance to ocean (ln)	-0.192***	-0.205***	-0.204***	-0.120***	-0.187***	-0.160***
Distance to nearest water body (ln)	-0.033***	-0.016***	-0.018***	-0.039***	-0.031***	-0.036***
Distance to nearest railroad (ln)	-0.007***	0.006***	0.008***	0.023***	0.004	0.006
Distance to nearest expressway (ln)	0.017***	0.027***	0.028***	0.029***	0.010***	0.021**
Airport noise > 60 LEQ	0.006	0.002	0.006	-0.149***	-0.016	-0.216**
School quality						
Elementary school						
A	0.106***	0.098***	0.089***	0.095***	0.122***	0.092**
B	-0.002	-0.002	-0.006	0.012	0.095***	0.093***
C	-0.059***	-0.051***	-0.046***	-0.018	0.056***	0.101***
Middle school						
A	0.238***	0.215***	0.193***	0.193***	0.050***	0.034
B	0.333***	0.321***	0.307***	0.269***	0.122***	0.062*
C	0.196***	0.155***	0.148***	0.128***	0.038***	0.000
High school						
B	0.275***	0.261***	0.253***	0.161***	0.047*	-0.029
C	0.264***	0.277***	0.267***	0.201***	0.002	-0.060
Month of sale						
February	0.008	0.009	0.010	0.011	0.009	0.010
March	0.029***	0.028**	0.028**	0.018*	0.024***	0.025***
April	0.020*	0.020*	0.021*	0.020***	0.025***	0.023***
May	0.048***	0.046***	0.046***	0.041***	0.046***	0.038***
June	0.047***	0.046***	0.046***	0.045***	0.047***	0.043***

Variable	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Month of sale						
July	0.049***	0.048***	0.049***	0.045***	0.052***	0.049***
August	0.055***	0.056***	0.055***	0.051***	0.055***	0.054***
September	0.048***	0.047***	0.048***	0.046***	0.052***	0.053***
October	0.050***	0.049***	0.049***	0.048***	0.055***	0.052***
November	0.056***	0.055***	0.056***	0.056***	0.065***	0.060***
December	0.058***	0.056***	0.056***	0.054***	0.062***	0.067***
Accessibility indexes (ln)						
Travel demand automobile	—	0.126***	—	—	—	—
Sugar Access automobile	—	—	0.175***	—	—	—
Distances to employment centers (ln)						
Nearest CBD	—	—	—	-0.061***	—	—
Nearest employment center	—	—	—	-0.183***	—	—
Submarket dummies	N	N	N	N	Y	N
Spatial lags						
Control variables	—	—	—	—	—	Y
Errors	—	—	—	—	—	0.921***
Adjusted R^2	*0.809*	*0.817*	*0.819*	*0.840*	*0.897*	*0.935*

Notes: *n*=13,932. *, **, and *** indicate significance at the 0.10, 0.05, and 0.01 levels, respectively. The table excludes estimates for the submarket dummy variables in Model 5 and the lagged control variables in Model 6.

Figure 1. Travel demand model and Sugar Access automobile accessibility indexes.



1a. Travel demand model index.

1b. Sugar Access index.