Do Atlanta residents value MARTA? Selecting an autoregressive model to recover willingness to pay.

Gregory S. Macfarlane[a], Laurie A. Garrow[a,*], Juan Moreno-Cruz[b]

[a] *School of Civil and Environmental Engineering, Georgia Institute of Technology, 790 Atlantic Drive, Atlanta GA 30332-0355*
[b] *School of Economics, Georgia Institute of Technology, 221 Bobby Dodd Way, Atlanta, GA 30332*

Abstract

Understanding homeowners' marginal willingness-to-pay (MWTP) for proximity to public transportation infrastructure is important for planning and policy. Naïve estimates of MWTP, however, may be biased as a result of spatial dependence, spatial correlation, and/or spatially endogenous variables. In this paper we discuss a class of spatial autoregressive models that control for these spatial effects, and apply them to sample data collected for the Atlanta, Georgia housing market. We provide evidence that a general-to-specific model selection methodology that relies on the generality of the spatial Durbin model (SDM) should be preferred to the classical specific-to-general methodology that begins with an assumption of no spatial effects. We show that applying the SDM raises the estimate of MWTP for transit proximity in Atlanta but also widens its confidence interval, relative to ordinary linear regression. This finding may have implications for risk estimations in land value capture forecasts and transportation policy decisions.

Keywords: spatial econometrics, spatial Durbin model, transportation accessibility, land value capture

1. Introduction

Home equity accounts for the largest share of household wealth in the United States, representing 78% of the net worth for the median household in 2010 (U.S. Census Bureau, 2010). Correspondingly, property taxes represent the largest independent revenue stream for local governments (Tax Policy Center, 2013). Municipal authorities have at least two interests in policy mechanisms that are correlated with raising home and property values: the welfare of their citizens and their municipalities' fiscal health.

One strategy that authorities may employ to raise property values is to construct transportation infrastructure. In theory, the rent price at locations with good accessibility to jobs, markets, schools or other activity centers will be higher than at locations with poor accessibility (Alonso, 1960). Advocates of public mass transit in particular point to a strategy of land value capture resulting

*Corresponding author. Tel.: +1.404.385.6634
Email addresses: gregmacfarlane@gatech.edu (Gregory S. Macfarlane), laurie.garrow@ce.gatech.edu (Laurie A. Garrow), juan.moreno-cruz@econ.gatech.edu (Juan Moreno-Cruz)

from transit development (Smith and Gihring, 2006): if a region[1] expends resources to improve the public transit system in a neighborhood, rents in that neighborhood should rise. This public expenditure will increase the private wealth of landholders in the improved area and consequently property tax revenues in the region. Whether government can recoup the cost of its infrastructure expenditure over a reasonable period of time, however, is an empirical question that remains unanswered, because the willingness of households to pay for a marginal improvement in their transportation accessibility is not entirely understood, and is perhaps heavily dependent on local circumstances.

The marginal willingness to pay (MWTP) for a characteristic of a good is a direct function of the utility derived from that characteristic (Rosen, 1974); therefore, a regression model with the price of the good as the dependent variable and the good's characteristics as predictor variables — called a *hedonic* model — reveals the MWTP for each characteristic if the assumptions of regression are met. The residual error terms, for instance, must be independently and identically distributed else the researcher cannot test that the MWTP is not zero. The challenge for researchers who develop hedonic home price models is that characteristics of urban housing markets interfere with regression assumptions in four important ways:

1. **Spatial dependence of prices:** The housing market is comparative; the price of a home is relative to the prices of homes nearby. This creates a missing variable bias in the linear regression model.
2. **Spatially correlated error terms:** Homes near each other have similar characteristics. This will violate the linear regression assumption that error terms are distributed independently, thus invalidating significance tests.
3. **Spatially endogenous or omitted variables:** Neighborhood attributes that are unobservable, such as neighborhood prestige, raise or lower the prices of homes. This can cause both a missing variable bias and cause correlated errors.
4. **Spatial heterogeneity (non-stationarity):** The housing market in one neighborhood may value some attributes more highly than the market in another. These differing preference are reflected in model parameters that vary in space.

Whereas the first three problems are types of *autocorrelation* and have a similar solution in autoregressive models, the solution to the fourth is to fit locally-weighted regressions as described by Brunsdon et al. (1999). This paper focuses on the first three problems and limits the discussion to autoregressive models. Spatial dependence and correlation are fundamentally different, although spatial endogeneity can be seen as a combination of both. Spatial dependence is a substantive problem, resulting in biased estimates of model parameters. Spatial correlation is a nuisance problem, affecting not the parameter estimates themselves but estimates of their standard errors. Identifying which problems may exist in a particular dataset or hedonic model is an important practical question for transportation researchers, on whose models transportation investment and policy decisions rely.

In this paper, we compare two modeling frameworks that have been used to identify spatial processes in housing markets with a particular emphasis on recovering the MWTP for public

[1] In the US, metropolitan planning organizations direct local, state, or federal funds to major transportation investments. Local governments generally collect property taxes. In a land value capture strategy, these disparate levels of government work in concert.

transit proximity. The first is the classical framework that relies on Lagrange multiplier tests for spatial dependence and correlation in the linear model residuals (Anselin, 1988b). The second is a general framework that tests for restrictions in a general nesting model (the spatial Durbin model, or SDM). The two frameworks may identify different preferred models in certain circumstances; previous comparative studies of hte frameworks using synthetic data (Florax et al., 2003; Larch and Walde, 2008) reached somewhat conflicting conclusions as a result of subtle distinctions in basic modeling assumptions. In spite of these subtleties,recent studies examining spatial autoregression in an accessibility context (Osland, 2010; Löchl and Axhausen, 2010; Ibeas et al., 2012) have used only the classical framework and may consequently have estimated biased parameters. We show that applying the general framework to home price data in Atlanta results in a materially different model than the classical framework. Further, our preferred model raises the expected MWTP relative to a simple linear model, but also expands the confidence interval around this estimate. This finding may have implications for risk estimations in land value capture strategies.

The remainder of this section provides the context for spatial econometric models in the larger hedonic evaluation literature. Section 2 reviews spatial econometric models and the two frameworks that have been used to select a preferred model. Sections 3 and 4 apply these two frameworks to the Atlanta region and the MARTA heavy-rail system and discuss results. Finally, Sections 5 and 6 compare our proposed framework to recent studies that have used one or both of the modeling frameworks and offer perspectives on future research objectives.

1.1. Home Prices and Transit Accessibility

Some of the earliest hedonic home price models showed a strong relationship between transportation network accessibility and home prices. Brigham (1965) showed that the accessibility potential of a parcel, defined by its access to highway networks, was a better predictor of home values in Los Angeles than its distance to the central business district. Dubin and Sung (1987) observed a similar result in Baltimore. Other early studies used highway accessibility as a control variable in a more holistic model of housing markets (Massell and Stewart, 1971; Ridker and Henning, 1967).

Researchers in the last thirty years have been particularly interested in the hedonic value of transit proximity, as many cities have opened or expanded public rail transit networks. Simple linear hedonic models abound, and generally show a positive MWTP for transit accessibility (Grass, 1992; Lewis-Workman and Brod, 1997). Other researchers have segmented their data or introduced variables to examine particular theories. Chen et al. (1998), for instance, showed that transit stations in Portland have both a positive proximity benefit on prices and also a negative nuisance effect stemming from the increase mechanical noise and foot traffic around station. Nelson (1992) observed two distinct MWTP estimates in Atlanta, with lower-income neighborhoods valuing transit proximity more than higher-income neighborhoods.

Urban housing markets are complicated systems, and researchers have applied numerous econometric techniques to isolate MWTP for transit proximity from other confounding variables. Bowes and Ihlanfeldt (2001) estimated sub-models of crime rate and retail activity in areas around Atlanta transit stations, and then used the predictions from these models as instruments in a hedonic model; this process may remove econometric endogeneity from variables that influence home prices but that are only indirectly related to the transit station. A number of studies have used time series or panel data methods to eliminate the effects of unobserved variables — with the assumption that these unobserved or endogenous variables do not change over time — and establish the direct effect of improved transportation accessibility on home prices (Chernobai et al., 2009; McMillen

and McDonald, 2004; Mikelbank, 2004; Iacono and Levinson, 2011). These methods may be less applicable to cities with mature transportation networks, where the transportation network is fixed and other variables in the housing market are changing instead.

The complexity of urban housing markets calls for econometric models that can provide unbiased estimates of MWTP for home or location characteristics, while still allowing for parsimonious model specification (Dubin et al., 1999). Numerous elements of a home or its neighborhood might influence its price, and to expect any researcher to capture all of these elements in a data vector is unrealistic. Rather than expand the variables included in an econometric model, the field of spatial econometrics leverages Tobler's first law of geography, that "nearer things are more related than distant things" (Tobler, 1970). By parsimoniously representing these relationships, spatial econometric models can produce unbiased measures of MWTP across a wide range of policy variables.

2. Methodology

The linear regression model, which is the traditional starting point for hedonic models, is expressed as

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{1}$$

where $\boldsymbol{y}$ is a vector of length n (the number of observations) and X is an $n \times p$ matrix of p attributes (including a constant intercept term). The average marginal effect of the attributes $\boldsymbol{x}_k \in X$ on $\boldsymbol{y}$ is given by the vector of slope parameters $\boldsymbol{\beta}$, which has an ordinary least-squares (OLS) estimator $\hat{\boldsymbol{\beta}}$

$$\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'\boldsymbol{y} \tag{2}$$

If the mean of the residuals $\boldsymbol{\epsilon}$ is 0, then the expected value of this estimator is unbiased. The variance of this estimator is $\sigma^2(X'X)^{-1}$, under the Gauss-Markov assumption that $\boldsymbol{\epsilon}$ is distributed independently and identically with a normal distribution of mean zero and variance σ^2; mathematically, $\mathrm{Var}\,(\boldsymbol{\epsilon}|X) = \sigma^2 I$. This assumption underlies all of the standard hypothesis tests on the significance of the elements of $\boldsymbol{\beta}$.

Two basic situations (among many) are known to interfere with OLS estimation. Variables that are excluded from X but which are nonetheless important to $\boldsymbol{y}$ will lead to biased estimates of $\boldsymbol{\beta}$. Error terms that are not independently or identically distributed will produce a biased estimate of σ, thereby invalidating hypothesis tests. Spatial dependence is effectively an omitted variable problem; a home's value depends at least partially on the values of nearby homes, and these values should therefore be incorporated into X. Spatial correlation, on the other hand, is econometrically similar to heteroskedasticity in that it creates unreliable estimates of model standard error.

2.1. Spatial Autoregressive Models

Spatial autocorrelation in a variable $\boldsymbol{x}$ may be represented by the relationship $\boldsymbol{x} = \rho W\boldsymbol{x}$, where ρ is a correlation coefficient and W is an $n \times n$ spatial weights matrix that maps each x_i onto its "neighbors" $x_j, j \in 1 \ldots n$. Elements w_{ij} of W are zero if i and j are not neighbors and positive if i and j are neighbors (more detail on spatial weights matrices is given in Dubin (1998). The correlation coefficient ρ is a measure of the strength of the spatial autocorrelation within $\boldsymbol{x}$; values of ρ close to zero indicate that there is little spatial autocorrelation in $\boldsymbol{x}$, and values close to one indicate that there is strong spatial autocorrelation in $\boldsymbol{x}$.

If the dependent variable $\boldsymbol{y}$ is autocorrelated, then the sample exhibits spatial dependence. A model that attempts to replicate this data generation process is the Cliff and Ord (1970) *spatial simultaneous autoregressive lag model* (SAR):

$$\begin{aligned} \boldsymbol{y} &= \rho W \boldsymbol{y} + X\boldsymbol{\beta} + \boldsymbol{\epsilon} \\ \boldsymbol{y} &= (I - \rho W)^{-1}(X\boldsymbol{\beta} + \boldsymbol{\epsilon}) \end{aligned} \tag{3}$$

This model, which contains a *spatial lag* of the dependent variable, will provide estimates of $\boldsymbol{\beta}$ that are robust to spatial dependence, provided that the researcher's specification of W is sufficiently close to the true, unobserved spatial structure. The corresponding model that addresses spatially autocorrelated errors is the *spatial error model* (SEM),

$$\begin{aligned} \boldsymbol{y} &= X\boldsymbol{\beta} + \boldsymbol{u}, \boldsymbol{u} = \lambda W \boldsymbol{u} + \boldsymbol{\epsilon} \\ \boldsymbol{y} &= X\boldsymbol{\beta} + (I - \lambda W)^{-1}\boldsymbol{\epsilon} \end{aligned} \tag{4}$$

Here λ is the correlation coefficient for the errors. Again, if W is sufficiently close to the true spatial relationship, the SEM will produce estimates of model standard error that are robust to residual autocorrelation. As the parameter $\lambda \geq 0$, the SEM must have *wider* standard errors than the OLS model.

A third model, the *spatial Durbin model* (SDM), was originally derived by Anselin (1980) as a consolidated form of the SEM

$$\boldsymbol{y} = \lambda W \boldsymbol{y} + X\boldsymbol{\beta} + WX(-\lambda\boldsymbol{\beta}) + \boldsymbol{\epsilon} \tag{5}$$

This model has spatial lags of both the dependent and the independent variables. The SDM may also be estimated in an unrestricted form by allowing the lagged independent variable parameter vector $-\lambda\boldsymbol{\beta}$ to take its own maximum likelihood value $\boldsymbol{\gamma}$ (Burridge, 1981).

$$\boldsymbol{y} = \rho W \boldsymbol{y} + X\boldsymbol{\beta} + WX\boldsymbol{\gamma} + \boldsymbol{\epsilon} \tag{6}$$

The unrestricted SDM is a linear combination of the SAR and SEM, and therefore is robust to both spatial dependence and spatial correlation.

The SDM has two further econometric properties that make it particularly appropriate for hedonic analysis. First, because the model contains a set of lagged predictors $\boldsymbol{\gamma}$, it explicitly models the externality that attributes of observation j impose on the outcome for observation i (Anselin, 2003). A consequence of this property is that the marginal effect of $\boldsymbol{x}_k$ on $\boldsymbol{y}$ is *not* β_k as in a linear model. Instead, there exist separate direct, indirect, and total effects that the analyst must consider. The theoretical average effects $M(k)$ of a variable $\boldsymbol{x}_k$ on $\boldsymbol{y}$ in a SDM are

$$\begin{aligned} M(k)_{\text{direct}} &= n^{-1} tr((I - \rho W)^{-1}(I\beta_k + W\gamma_k)) \\ M(k)_{\text{total}} &= n^{-1}\iota'(I - \rho W)^{-1}(I\beta_k + W\gamma_k)\iota \\ M(k)_{\text{indirect}} &= M(k)_{\text{total}} - M(k)_{\text{direct}} \end{aligned} \tag{7}$$

where ι is a vector of ones of length n. It is important to note that all three types of effects are linear functions of β_k, γ_k, W, and ρ. Details on efficiently calculating these effects are given by LeSage and Pace (2009, p. 114). Concisely, a Monte Carlo simulation of the effects based on draws of β_k, γ_k, and ρ based on the analytical model parameter variance-covariance matrix can produce empirical standard errors of the effects.

This explicit modeling of direct and indirect effects has made the SDM popular in modeling systems where externalities are important, such as the home price penalty of being downwind from swine farms (Kim and Goldsmith, 2008). The SDM is also especially appropriate for systems where the observations interact with each other, such as trade between regions (LeSage and Fischer, 2008). Whether the analyst pays more attention to direct, indirect, or total effects will depend on her specific problem; recommended policy interventions are dependent on which type of effect is considered. In the case of MWTP for transit accessibility, we are mostly concerned with the total effect.

The second econometric property of the SDM is that it may be seen to arise from an autoregressive fixed effects process, and can therefore control for spatially correlated endogenous or omitted variables, given some assumptions. Neighborhood prestige, for example, is not an attribute that can be measured properly; school quality or crime rates may play a role, but these variables are endogenous in that quality neighborhoods create quality schools, and vice-versa. Assume that each observation i inherits some unobserved fixed effects based on its spatial location a_i. We wish to include the entire vector of fixed effects $\boldsymbol{a}$ (a vector of length n) in a model to control for the endogenous omitted variables and to remove all correlation between X and ε, but such a model would be inestimable as it would contain $n + p$ variables and only n observations. If we assume, however, that the fixed effects $\boldsymbol{a}$ follow a spatial autoregressive process and are correlated with the X terms, we can construct the data generating process

$$\boldsymbol{a} = \rho W \boldsymbol{a} + X\boldsymbol{\gamma}' + \boldsymbol{\epsilon} \tag{8}$$

with ρ a correlation coefficient, $\boldsymbol{\gamma}'$ a vector of estimable parameters of length p, and $\boldsymbol{\epsilon}$ assumed to be distributed IID normal. Solving Equation 8 for $\boldsymbol{a}$ yields an expression for the fixed effects

$$\boldsymbol{a} = (I - \rho W)^{-1}(X\boldsymbol{\gamma}' + \boldsymbol{\epsilon}) \tag{9}$$

Replacing the error of the linear model with the spatial autoregressive fixed effect given in Equation 9 and rearranging terms,

$$\begin{aligned}
\boldsymbol{y} =& X\boldsymbol{\beta} + (I - \rho W)^{-1}(X\boldsymbol{\gamma}' + \boldsymbol{\epsilon}) \\
(I - \rho W)\boldsymbol{y} =& (I - \rho W)X\boldsymbol{\beta} + X\boldsymbol{\gamma}' + \boldsymbol{\epsilon} \\
\boldsymbol{y} =& \rho W\boldsymbol{y} + X(\boldsymbol{\beta} + \boldsymbol{\gamma}') + WX(-\rho\boldsymbol{\beta}) + \boldsymbol{\epsilon} \\
\boldsymbol{y} =& \rho W\boldsymbol{y} + X\boldsymbol{\beta}^{*} + WX\boldsymbol{\gamma} + \boldsymbol{\epsilon}
\end{aligned}$$

gives the SDM presented in Equation 5 (LeSage and Pace, 2009, p. 29). This analysis implies that the effects of neighborhood variables that are omitted or unobservable, such as school quality, crime rates, or neighborhood prestige are controlled for with the stipulation that these variables themselves follow a spatial autoregressive process. If, on the other hand, the missing variables are spatially uncorrelated or exogenous, then spatial models may be unnecessary.

2.2. Model Selection

Table 1 presents a summary of the consequences for using a mis-specified spatial model. A failure to account for spatial dependence, by using an OLS or SEM when an SAR or SDM is the true model, results in biased estimates of the model parameters. Failure to account for correlated model errors, by using an OLS or SAR when an SEM or SDM is the true model, will result in

Table 1: Consequences of misspecified hedonic model.

	Estimated Model			
True DGP	OLS	SAR	SEM	SDM
OLS: $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$	-	inefficient	inefficient	inefficient
SAR: $\boldsymbol{y} = \rho W\boldsymbol{y} + X\boldsymbol{\beta} + \boldsymbol{\epsilon}$	$\hat{\boldsymbol{\beta}}$ biased	-	$\hat{\boldsymbol{\beta}}$ biased	inefficient
SEM:$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \boldsymbol{\epsilon} = \lambda W\boldsymbol{\epsilon} + \boldsymbol{u}$	$\hat{\sigma}^2$ invalid	$\hat{\sigma}^2$ invalid	-	inefficient
SDM: $\boldsymbol{y} = \rho W\boldsymbol{y} + X\boldsymbol{\beta} + WX\boldsymbol{\gamma} + \boldsymbol{u}$	$\hat{\boldsymbol{\beta}}$ biased	$\hat{\sigma}^2$ invalid	$\hat{\boldsymbol{\beta}}$ biased	-

biased estimates of standard errors and the invalidation of parameter significance tests. Using any spatial model when it is not required reduces the efficiency of the estimates, as the analyst will sacrifice degrees of freedom to estimate unneeded parameters. Testing whether ρ or λ are zero in the estimates of Equation 3 or Equation 4 is direct but inadequate, as spatial correlation may appear to be spatial dependence and vice versa; a more robust selection framework is required. There are two primary frameworks that analysts may use to select the appropriate spatial autoregressive model: a classical framework that uses forward-looking statistics based on OLS residuals, and a general framework that compares autoregressive models with likelihood ratio tests.

2.2.1. Classical Framework

Least-squares estimates of spatial autoregressive models are inconsistent and the models must therefore be estimated using maximum likelihood techniques. The log-likelihood function for the SAR model (for example) is (Anselin, 1988b):

$$\ln(\mathcal{L}_{SAR}) = -(n/2)\ln(\pi\sigma^2) + \ln|I - \rho W| - \frac{e'e}{2\sigma^2}, \tag{10}$$
$$e = \boldsymbol{y} - \rho W\boldsymbol{y} - X\boldsymbol{\beta}$$

Calculating the determinant of an arbitrary $n \times n$ matrix is a computationally expensive process of $O(n!)$, and the log-determinant term $\ln|1 - \rho W|$ is present in the likelihood functions for the SAR, the SEM, and the SDM models. There are features of W that reduce the computational order LeSage and Pace (2009, Chap. 4), but even on a modern computer this is a time-consuming process. When these models were first developed, there was an incentive to find tests for autocorrelation that avoided computing the likelihood function in Equation 10. The classical framework was born from the need to avoid intensive computer calculations.

Anselin (1988a) developed Lagrange multiplier tests for spatial dependence (LM_ρ) and spatial correlation (LM_λ) that are estimated using the OLS residuals, $\hat{\boldsymbol{\epsilon}} = \boldsymbol{y} - X\hat{\boldsymbol{\beta}}$:

$$LM_\rho = \frac{(\hat{\boldsymbol{\epsilon}}' W\boldsymbol{y}/\hat{\sigma}^2)^2}{nJ} \tag{11}$$

$$LM_\lambda = \frac{(\hat{\boldsymbol{\epsilon}}' W\hat{\boldsymbol{\epsilon}}/\hat{\sigma}^2)^2}{T} \tag{12}$$

with

$$J = \frac{1}{n\hat{\sigma}^2}[(WX\hat{\boldsymbol{\beta}})'(I - X(X'X)^{-1}X')(WX\hat{\boldsymbol{\beta}}) + T\hat{\sigma}^2] \tag{13}$$

and T the trace of the matrix $W'W+W^2$. Both of these statistics are asymptotically distributed χ^2_1. Rejecting the null hypothesis that $LM_\rho = 0$ implies that the proper model is the SAR. Similarly, rejecting that $LM_\lambda = 0$ implies that the proper model is the SEM.

The LM tests suffer from the same confusion as the direct parametric tests on λ and ρ; that is, spatial dependence can appear as spatial correlation in the model residuals and vice versa. It is therefore not uncommon that *both* LM tests will reject the null hypothesis. For this reason, Anselin et al. (1996) proposed *robust* LM tests:

$$RLM_\rho = \frac{(\hat{\boldsymbol{\epsilon}}'W\hat{\boldsymbol{\epsilon}} - T(nJ)^{-1}\hat{\boldsymbol{\epsilon}}'W\boldsymbol{y}/\hat{\sigma}^2)^2}{nJ - T} \tag{14}$$

$$RLM_\lambda = \frac{(\hat{\boldsymbol{\epsilon}}'W\boldsymbol{y} - \hat{\boldsymbol{\epsilon}}'W\hat{\boldsymbol{\epsilon}}/\hat{\sigma}^2)^2}{T[1 - T(nJ)]^{-1}} \tag{15}$$

It is even possible that both of the RLM statistics will reject their null hypotheses; in this case, Florax et al. (2003) recommend selecting the model (either SAR or SEM) with the larger test statistic. This framework does not lead to the unrestricted SDM model, though in the presentation by Osland (2010), the SDM should be used if the robust LM tests are "inconclusive." This selection framework is described in Algorithm 1.

2.2.2. General Framework

An alternative strategy that we term the "general" framework, was proposed by Florax et al. (2003), and relies on the fact that the SDM is a linear combination of the SAR and SEM. Previously, Hendry (1979) proposed that whenever a general nesting model (such as the SDM) exists, it is appropriate to begin the specification search there. LeSage and Pace (2009) argue that from a Bayesian model uncertainty perspective, this alternative strategy is the only appropriate approach (page 31). Consider the SDM ($\boldsymbol{y}_c$) as a weighted linear combination of the SAR ($\boldsymbol{y}_a$) and SEM ($\boldsymbol{y}_b$) models,

$$\begin{aligned}
\boldsymbol{y}_c &= \pi_a\boldsymbol{y}_a + \pi_b\boldsymbol{y}_b \\
\boldsymbol{y}_c &= \pi_a((I-\rho W)^{-1}(X\boldsymbol{\beta} + \boldsymbol{\epsilon})) + \pi_b(X\beta + (I-\rho W)^{-1}\boldsymbol{\epsilon}) \\
(I-\rho W)\boldsymbol{y}_c &= X(\pi_a\boldsymbol{\beta}) + (I-\rho W)(X\pi_b\boldsymbol{\beta}) + (\pi_a+\pi_b)\boldsymbol{\epsilon} \qquad (16)\\
\boldsymbol{y}_c &= \rho W\boldsymbol{y_c} + (\pi_a+\pi_b)X\boldsymbol{\beta} + WX(-\rho\pi_b\boldsymbol{\beta}) + \boldsymbol{\epsilon} \qquad (17)
\end{aligned}$$

with π_a the probability of an SAR and π_b the probability of an SEM, $\pi_a + \pi_b = 1$. If the true specification is an SAR, then no data evidence will show that $\pi_b > 0$, and Equation 16 will reduce to the SAR. Conversely, if the true specification is an SEM, then $\pi_a = 0$, and the SEM will remain. For any situation in which there exists uncertainty, $0 < \pi_a, \pi_b < 1$, the full SDM in Equation 17 should be used.

The analyst implements this framework by estimating the SDM and SEM models. If the true model is the SEM, then $\boldsymbol{\gamma} = -\rho\boldsymbol{\beta}$ (from Equation 6), and the two models will have the same model likelihood. A likelihood ratio (LR) test can be used,

$$-2(\ln(\mathcal{L}_{SEM}) - \ln(\mathcal{L}_{SDM})) \sim \chi^2_1 \tag{18}$$

If, on the other hand, the true model is an SAR, then the lagged independent parameters $\boldsymbol{\gamma} = 0$ and the reduction is trivial. Finally, if $\rho = 0$, then an OLS model is sufficient. Details of this selection framework are given in Algorithm 2.

Algorithm 1 Classical Selection Framework

procedure SpEffects($\boldsymbol{y}, X, W$)
  Obtain $\hat{\boldsymbol{\epsilon}} = \boldsymbol{y} - X(X'X)^{-1}X'\boldsymbol{y}$ ▷ OLS residuals
  Calculate $LM_\rho : \rho \stackrel{?}{=} 0$ AND $LM_\lambda : \lambda \stackrel{?}{=} 0$ ▷ Lagrange multiplier tests
  if $\rho = 0$ AND $\lambda = 0$ **then**
    OLS: $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$ ▷ Efficient, risk bias in $\boldsymbol{\beta}, \sigma$
  else if $\rho \neq 0$ AND $\lambda = 0$ **then**
    SAR: $\boldsymbol{y} = \rho W\boldsymbol{y} + X\boldsymbol{\beta} + \boldsymbol{\epsilon}$ ▷ $\boldsymbol{\beta}$ unbiased, risk in σ
  else if $\rho = 0$ AND $\lambda \neq 0$ **then**
    SEM: $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \boldsymbol{\epsilon} = \lambda W\boldsymbol{\epsilon} + \boldsymbol{u}$ ▷ σ unbiased, risk in $\boldsymbol{\beta}$
  else
    Calculate $RLM_\rho : \rho \stackrel{?}{=} 0$ AND $RLM_\lambda : \lambda \stackrel{?}{=} 0$ ▷ Robust LM tests
    if $\rho \neq 0$ AND $\lambda = 0$ **then**
      SAR: $\boldsymbol{y} = \rho W\boldsymbol{y} + X\boldsymbol{\beta} + \boldsymbol{\epsilon}$ ▷ $\boldsymbol{\beta}$ unbiased, risk in σ
    else if $\rho = 0$ AND $\lambda \neq 0$ **then**
      SEM: $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \boldsymbol{\epsilon} = \lambda W\boldsymbol{\epsilon} + \boldsymbol{u}$ ▷ σ unbiased, risk in $\boldsymbol{\beta}$
    else if $\rho \neq 0$ AND $\lambda \neq 0$ **then**
      if $RLM_\rho > RLM_\lambda$ **then**
        SAR: $\boldsymbol{y} = \rho W\boldsymbol{y} + X\boldsymbol{\beta} + \boldsymbol{\epsilon}$ ▷ $\boldsymbol{\beta}$ unbiased, risk in σ
      else if $RLM_\lambda > RLM_\rho$ **then**
        SEM: $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \boldsymbol{\epsilon} = \lambda W\boldsymbol{\epsilon} + \boldsymbol{u}$ ▷ σ unbiased, risk in $\boldsymbol{\beta}$
      else
        SDM: $\boldsymbol{y} = \rho W\boldsymbol{y} + X\boldsymbol{\beta} + WX\boldsymbol{\gamma} + \boldsymbol{\epsilon}$ ▷ $\boldsymbol{\beta}, \sigma$ unbiased, risk inefficiency
      end if
    end if
  end if
end procedure

Algorithm 2 General Selection Framework

procedure SpEffects($\boldsymbol{y}, X, W$)
  Estimate SDM: $\boldsymbol{y} = \rho W\boldsymbol{y} + X\boldsymbol{\beta} + WX\boldsymbol{\gamma} + \boldsymbol{\epsilon}$ ▷ $\boldsymbol{\beta}, \sigma$ unbiased, risk inefficiency
  if $\boldsymbol{\gamma} = -\rho\boldsymbol{\beta}$ **then**
    SEM:$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \boldsymbol{\epsilon} = \lambda W\boldsymbol{\epsilon} + \boldsymbol{u}$ ▷ σ unbiased, risk in $\boldsymbol{\beta}$
    if $\lambda = 0$ **then** ▷ No correlation
      OLS: $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$ ▷ Efficient, risk bias in $\boldsymbol{\beta}, \sigma$
    end if
  else if $\boldsymbol{\gamma} = \mathbf{0}$ **then**
    SAR: $\boldsymbol{y} = \rho W\boldsymbol{y} + X\boldsymbol{\beta} + \boldsymbol{\epsilon}$ ▷ $\boldsymbol{\beta}$ unbiased, risk in σ
    if $\rho = 0$ **then** ▷ No Dependence
      OLS: $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$ ▷ Efficient, risk bias in $\boldsymbol{\beta}, \sigma$
    end if
  else
    SDM: $\boldsymbol{y} = \rho W\boldsymbol{y} + X\boldsymbol{\beta} + WX\boldsymbol{\gamma} + \boldsymbol{\epsilon}$
  end if
end procedure

2.2.3. Comparisons of the Two Frameworks

There have been a number of analytical studies comparing the two modeling frameworks. In essentially all of these studies, the researcher creates a known data generating process (DGP), and executes a Monte Carlo simulation estimating models to recover this DGP. In the previously cited work by Florax et al. (2003), the researchers find that the classical approach identifies the correct DGP (which is an SAR, SEM, or OLS model) more frequently than the LR-based general approach. For high values of spatial correlation however (ρ, λ approaching 1), the two methods produce similar outcomes. The Florax et al. (2003) study did not consider that the true model might be an SDM.

By contrast, Larch and Walde (2008) showed that when the SDM was an available option the general framework is preferred. Further, they recommend that Wald tests for common factors be used instead of LR tests, because their power was higher in the range of autocorrelation tested. The true spatial parameters used by Larch and Walde were very small, however, with $\rho, \lambda \leq 0.2$. This corroborates the results of Mur and Angulo (2006), who compared the power of the LR test statistic for common factors against LM and Wald tests of the same hypothesis. They showed that, with large sample sizes and high R^2 statistics in the OLS model, that the three tests performed equivalently. This implies that the choice of statistic depends on the strength of spatial correlation, but that at higher levels the difference is negligible.

In summary, previous analytical evidence suggests that for systems where the SDM is a candidate model and where spatial dependence is likely, the general framework using likelihood ratio statistics is preferred.

3. Empirical Application

3.1. Data

This study used targeted marketing (TM) records, which are maintained by credit reporting agencies and other private firms in an effort to assess the creditworthiness of adults in the U.S. TM

records are compiled from bank records, credit card statements, and other sources both public and private, and contain detailed information on household demographics, financial obligations, and consumption patterns. TM records represent an emerging and potentially important source of data for transportation research (Kressner and Garrow, 2012). A sample of TM records representing the 13-county Atlanta metropolitan area was joined to transportation network shapefiles available from the Atlanta Regional Commission (2012). We restricted the analysis to a cross-section of owner-occupied homes purchased in 2009 and 2010 located within five miles of a Metropolitan Atlanta Rapid Transit Authority (MARTA) heavy rail transit station. Using a cross-section of sales from a small time frame eliminates the need to consider temporal dependence; spatial autoregressive models can accommodate temporal effects, but LeSage and Pace (2009, chap. 2) also show that a cross-sectional spatial autoregressive model represents the equilibrium point of a temporally dependent process. The MARTA system operates in only two of metropolitan Atlanta's 13 counties; limiting the scope of the analysis to homes within five miles of a MARTA station avoids confusing proximity to MARTA with proximity to central Atlanta, two measurements which will be highly collinear for homes near the periphery of the region.

The 4,812 observations, mapped in Figure 1, match our expectation of settlement patterns in the city. In the northern part of the study area a number of observations are missing in a shape corresponding to the boundaries of the City of Sandy Springs. The similarly large empty space in the south corresponds to a military base and Atlanta Hartsfield-Jackson International Airport. Descriptive statistics for the sample are given in Table 2. Analysis available from the authors shows that this sample does not deviate materially from data for the Atlanta region collected through the US Census Bureau's American Community Survey.

3.2. Model

We predict the price of a home as a function of its proximity to a rail station, conditioned on attributes of the home. The basic regression model is semi-log,

$$\log(\mathit{Value})_i = f(\boldsymbol{H}_i, \boldsymbol{O}_i, \boldsymbol{A}_i) \tag{19}$$

where $\boldsymbol{H}_i$ is a vector of house attributes including the square footage and property acreage of a home, its structural type (single dwelling, multiple dwellings, or mobile), and its age. $\boldsymbol{O}_i$ is a vector of attributes describing the homeowner, namely the household annual income and the ethnicity of the householder.

$\boldsymbol{A}_i$ describes the transportation accessibility of the home, which for our study is the natural logarithm of the Euclidean distance (in miles) between the home and its nearest MARTA station. The existing literature on hedonic models of transportation accessibility uses a wide array of specifications, including linear distance (Grass, 1992), binary proximity (Bowes and Ihlanfeldt, 2001), linear distance conditioned on binary proximity (Hess and Almeida, 2007), and spline regression (Chernobai et al., 2009). Proximity is an inherently continuous phenomenon: a home 0.51 miles from a rail station is only 100 feet from a home that is 0.49 miles from the station, a negligible distance for virtually all homebuyers. Further, some homebuyers may feel they have access at 0.75 miles, and others feel they do not past 0.25 miles. This fact is highlighted by Debrezion et al. (2007), who find continuous functions generally are better predictors of residential property values than proximity dummies (though they find the reverse is true for commercial properties). Using the natural logarithm applies a diminishing marginal cost to the distance; that is, 100 feet proportionally adds more to the cost of a 400-foot journey than to a journey of 2,000 feet. This

Table 2: Descriptive statistics of model variables

Continuous Variables	Mean	Median	Std. Dev.	Min	Max
Market value of home (kUSD)	311	225	274	50	2,000
Home built area (square feet)	2,125	1,875	1,115	750	7,000
Property area (acres)	0.425	0.375	1.22	0.25	40
Age of home (years)	38.2	38	24.8	1	120
Household Income (kUSD)	84.5	62.5	56.8	10	250
Distance to MARTA rail station (miles)	2.24	2.03	1.39	0.0403	5.21
Distance to freeway entrance (miles)	1.4	1.26	0.835	0.0767	4.61

Discrete Variables	Number	%
Property type		
Single dwelling unit	4,318	87.8
Multiple dwelling units (condominium)	599	12.2
Ethnicity		
White	2,644	53.8
African-American	1,567	31.9
Asian	175	3.56
Hispanic	144	2.93
Other	387	7.87

same logic of diminishing marginal cost or returns compels us to similarly log-transform the home value, household income, square footage, and property acreage for the observations. Highway accessibility also should affect home prices; we control for this by including the natural logarithm of the distance in miles to the nearest freeway entrance point.

The goal of this research is to highlight how much estimates of MWTP for transit accessibility can vary based on the spatial model selection methodology used by the researcher, and not to provide a conclusive measurement of MWTP for transit in Atlanta. We believe that we have selected a plausible model specification to fit the data we possess, but acknowledge that there may be other variables that could prove significant, improve the model fit, and/or alter the range of MWTP returned by each model.

3.3. Spatial Weights

Selecting an appropriate spatial weights matrix is essential, as misspecification itself may be a source of bias in the econometric model (Páez et al., 2008). However, the analyst must generally specify the matrices *a priori*, with little information on the appropriate form (Dubin, 1998). For the autoregressive models in this study, houses were neighbors if they were located within 1.8 miles of each other. The link was weighted by the inverse distance between neighbors to give more consideration to nearer observations. Our neighbors matrix W is the row-standardized inverse Euclidean distance between observations

$$W_{ij} = \begin{cases} \frac{1/d_{ij}}{\sum_{k=1}^{n}(1/d_{ik})} & \text{for } d_{ij} \leq 1.8 \text{ miles} \\ 0 & \text{for } d_{ij} > 1.8 \text{ miles} \end{cases} \tag{20}$$

Row-standardization is a technique that aids in identification and interpretation. We selected this weighting scheme and radius because it returned the highest model likelihood in a comparison of 45 different schema including nearest neighbors, unweighted neighbors, and contiguous polygons of varying distances and orders of adjacency balanced against the desire for conservative coefficients. The full details of this selection process are given in Appendix B.

4. Model results

Maximum likelihood estimates of the OLS, SAR, SEM, and SDM models were calculated using the "spdep" package for R (Bivand, 2006; R Development Core Team, 2013); the parameter estimates and statistics are given in Table 3.

Classical Selection. The results of the LM tests are presented in Table 4. As shown, we reject the null hypothesis that there is no spatial dependence or spatial correlation in the model. As the RLM_λ statistic is an order of magnitude larger than the RLM_ρ statistic, the SEM is conclusively the preferred model.

In the classical selection procedure, we follow Algorithm 1. Using the OLS residuals and the trace of the spatial weights matrix, we compute the LM tests for spatial effects; the test statistics are presented in Table 4. As both LM_ρ and LM_λ reject the null hypothesis that there is no dependence or correlation, we proceed to line 10. We calculate the robust LM test statistics (also in Table 4). As before, we reject both null hypotheses that there is no spatial dependence or spatial correlation in the model, and proceed to line 16. As the RLM_λ statistic is an order of magnitude larger than the RLM_ρ statistic, we move to line 20, and select the SEM as the preferred model.

General Selection. In the general selection procedure, we follow Algorithm 2. We estimate the SDM model in its restricted ($y = -\rho\vec{\beta}$) and unrestricted form. The LR test statistic produces a p-value less than 1×10^{-16}, so we reject that the two models are the same, and therefore infer $y \neq -\rho\vec{\beta}$. We can also see in the unrestricted SDM that not all $_k$ parameters are 0, so we proceed to line 14 and select the SDM as the preferred model.

This disagreement between the two selection frameworks methodologies is concerning, and suggests that there may be differences in model interpretation or inference.

4.1. Interpretation of Model Coefficients

The estimated effects of the model variables on home price are given in Table 5; the distribution of these effects is calculated empirically with repeated draws from the parameter variance-covariance matrices. The added information gained from modeling spatial dependence can be seen by examining the effects of home age on home price in detail. In the OLS model, a one-year increase in the age of a home cannot be said to have any relationship on its value. After controlling for correlated errors (with the SEM), an additional year is measured to have a -0.275 percent impact, a small estimate with a high degree of statistical significance. Controlling for spatial dependence (with the SAR) similarly shows negative direct, indirect, and total effects of an additional year in age. But in an SAR, the direct and indirect effects (and consequentially, the total effects) are required to have the same sign, a condition relaxed in an SDM. Indeed, the SDM shows a significant relationship between a homes age and its value but the direct and indirect effects conflict, with a direct effect of -0.003 and an indirect effect of 0.012. This is intuitive: living in an older home for

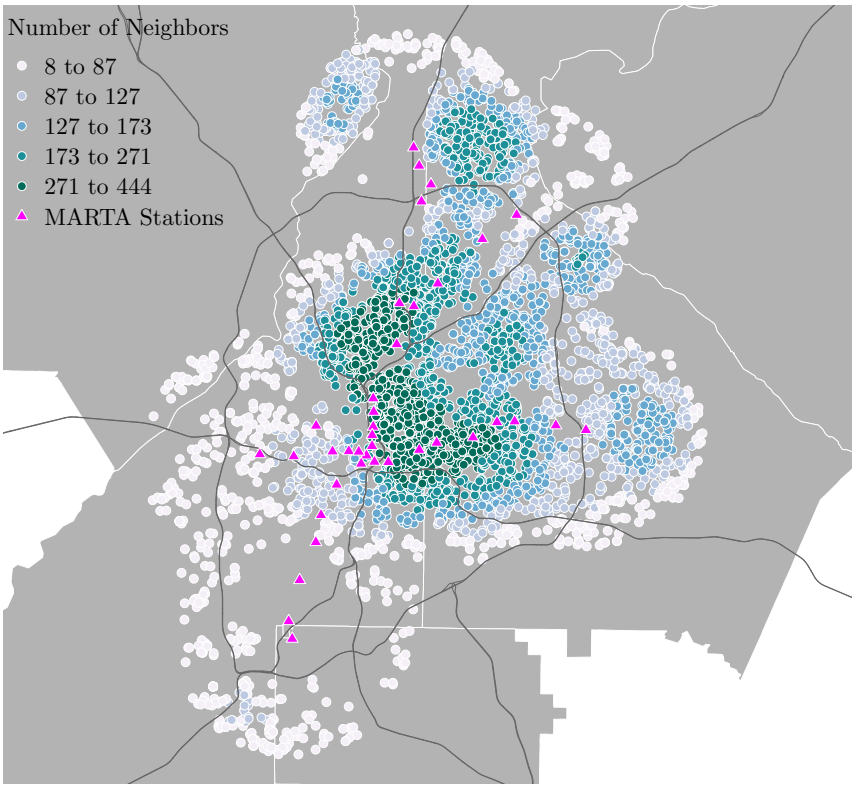

Figure 1: Observations by number of neighbors.

Table 3: Estimated model parameters and statistics.

	OLS		SAR		SEM		SDM			
Covariate	β	t-stat	β	t-stat	β	t-stat	β	t-stat	γ	t-stat
ρ			0.677^{***}	67.5			0.859^{***}	51.8		
λ					0.975^{***}	171.6				
(Intercept)	-1.866^{***}	-13.9	3.506^{***}	-32.6	-0.959^{***}	-4.9	-1.431^{***}	-5.3		
Property type: *ref. Single unit*										
Multiple Units	-0.043^{*}	-2.1	-0.139^{***}	-8.6	-0.312^{***}	-17.2	-0.319^{***}	-17.4	0.386^{***}	10.0
Home age	0.000	1.3	-0.001^{***}	-3.5	-0.003^{***}	-14.3	-0.003^{***}	-14.8	0.005^{***}	9.0
log(Square feet)	0.915^{***}	57.0	0.731^{***}	56.1	0.783^{***}	64.5	0.776^{***}	63.7	-0.521^{***}	-12.0
log(Lot acres)	0.046^{***}	3.8	0.025^{**}	2.6	0.090^{***}	9.7	0.091^{***}	9.8	-0.139^{***}	-5.1
Race: *ref. White*										
African-American	-0.295^{***}	-21.6	-0.044^{***}	-3.9	-0.042^{***}	-3.4	-0.038^{**}	-3.1	0.021	0.9
Hispanic	-0.105^{***}	-3.3	-0.008	-0.3	-0.029	-1.4	-0.027	-1.2	0.009	0.1
log(Income)	0.263^{***}	30.5	0.065^{***}	8.8	0.075^{***}	9.5	0.070^{***}	8.9	-0.011	-0.6
log(Distance from MARTA)	-0.192^{***}	-25.2	-0.142^{***}	-23.8	-0.030^{*}	-2.2	-0.006	-0.4	-0.027	-1.4
log(Distance from freeway entrance)	-0.047^{***}	-5.4	-0.041^{***}	-5.9	0.028^{*}	2.3	$0.026^{\dagger}$	1.9	$-0.034^{\dagger}$	-1.7
N	4,917		4,917		4,917		4,917			
$\log(\mathcal{L})$	$-2,306$		-875		-331		-271			

† significant at $p < .10$; $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$

Table 4: Lagrange multiplier tests for spatial effects.

Test	Statistic	p-value
LM_ρ	3,057	$< 1 \times 10^{-16}$
LM_λ	8,690	$< 1 \times 10^{-16}$
RLM_ρ	531	$< 1 \times 10^{-16}$
LM_λ	6,614	$< 1 \times 10^{-16}$

its own sake brings few benefits, but living around older homes might be a sign of situation in a more established (or even historic) neighborhood.

Similar logic can be applied to interpreting the effects of multiple unit dwellings: OLS, SAR, and SEM all suggest that the market values condominiums less than detached homes, but the SDM suggests that nearby condominiums contribute value. This, again, is intuitive. Isolated condominiums in neighborhoods where single-family homes are typical will be seen as less attractive. On the other hand, condominiums dominate the housing market in some of Atlantas most exclusive neighborhoods, and nearby homes are the more valuable for it. The direct and indirect effects of income, however, are both positive: as the income of a homeowner rises or the incomes of the neighbors the value of the home will rise also.

Other variables show similar trends across the models. Take ethnicity, for example. In the OLS model, homes with African-American and Hispanic residents are valued significantly less than homes with White residents. In a way this is a strange observation, because when someone buys a home the existing residents usually do not remain and should not be considered in the price. In the SDM model, both the direct and indirect effects of Hispanic owners and neighbors lose significance. Further, the direct effect of African-American owners has been reduced tenfold, suggesting that most of the observed OLS effect might actually belong to variables that are spatially correlated with the location of African-American homes, but that African-Americans do not themselves substantially lower home prices.[2]

Some of these relationships could be uncovered with OLS model specifications that were carefully tailored to look at these particular issues. Studies that attempt to examine MWTP for a specific variable may want to take care that the variables of interest are properly represented, but the SDM successfully controls for these relationships with a parsimonious model (Dubin et al., 1999).

4.2. Interpretation of MWTP for Transit Accessibility

Our variable of primary interest, the "Miles from MARTA" variable, is significant at the 95% confidence level in three of the four models (excluding the SEM), though the direct effect in the SDM is not significant (in this specification, a more negative value represents a higher MWTP). As discussed in Section 2, the total effect is the most appropriate measurement for MWTP for transit accessibility. This is because it is difficult to conceive a scenario where transit is "moved" closer to a home without also moving closer to the home's neighbors. Using the SDM model, we estimate that a 1% increase in the distance between a home and its nearest transit station lowers the expected value of the home by 0.235 percent, all else equal. Figure 2 shows 95% confidence bands of our estimated MWTP for transit proximity. As per the discussion in Section 2, we expect

[2]Though the effect is significant, it is not large enough to be meaningful.

Table 5: Average marginal effect of model variables.

	OLS		SAR		SEM		SDM	
Covariates	Effect	t-stat	Effect	t-stat	Effect	t-stat	Effect	t-stat
Direct Effects								
Multiple Units	−0.043	−2.1	−0.142	−8.6	−0.312	−17.2	−0.312	−17.3
Home age	0.000	1.27	−0.001	−3.49	−0.003	−14.3	−0.003	−14.3
log(Square feet)	0.915	57	0.747	56.9	0.783	64.5	0.785	64.7
log(Lot acres)	0.046	3.83	0.025	2.6	0.090	9.73	0.087	9.47
African-American	−0.295	−21.6	−0.045	−3.9	−0.042	−3.35	−0.0389	−3.13
Hispanic	−0.105	−3.33	−0.009	−0.337	−0.029	−1.36	−0.027	−1.17
log(Income)	0.263	30.5	0.066	8.86	0.075	9.53	0.073	9.36
log(Distance from MARTA)	−0.192	−25.2	−0.146	−23.8	−0.030	−2.17	−0.008	−0.568
log(Distance from freeway entrance)	−0.047	−5.37	−0.042	−5.92	0.028	2.32	0.025	1.88
Indirect Effects								
Multiple Units			−0.288	−8.16			0.789	3.45
Home age			−0.001	−3.47			0.015	4.8
log(Square feet)			1.51	22.4			0.999	4.18
log(Lot acres)			0.051	2.61			−0.429	−2.43
African-American			−0.091	−4.06			0.080	−0.587
Hispanic			−0.017	−0.334			−0.092	−0.137
log(Income)			0.134	9.56			0.345	3.37
log(Distance from MARTA)			−0.295	−17.2			−0.227	−3.43
log(Distance from freeway entrance)			−0.0844	−5.59			−0.0848	−1.1
Total Effects								
Multiple Units			−0.43	−8.42			0.477	2.09
Home age			−0.002	−3.49			0.012	3.93
log(Square feet)			2.26	31.8			1.78	7.42
log(Lot acres)			0.077	2.61			−0.342	−1.93
African-American			−0.136	−4.02			−0.119	−0.868
Hispanic			−0.026	−0.335			−0.12	−0.175
log(Income)			0.2	9.5			0.418	4.16
log(Distance from MARTA)			−0.441	−20.4			−0.235	−3.89
log(Distance from freeway entrance)			−0.441	−20.4			−0.060	−0.836

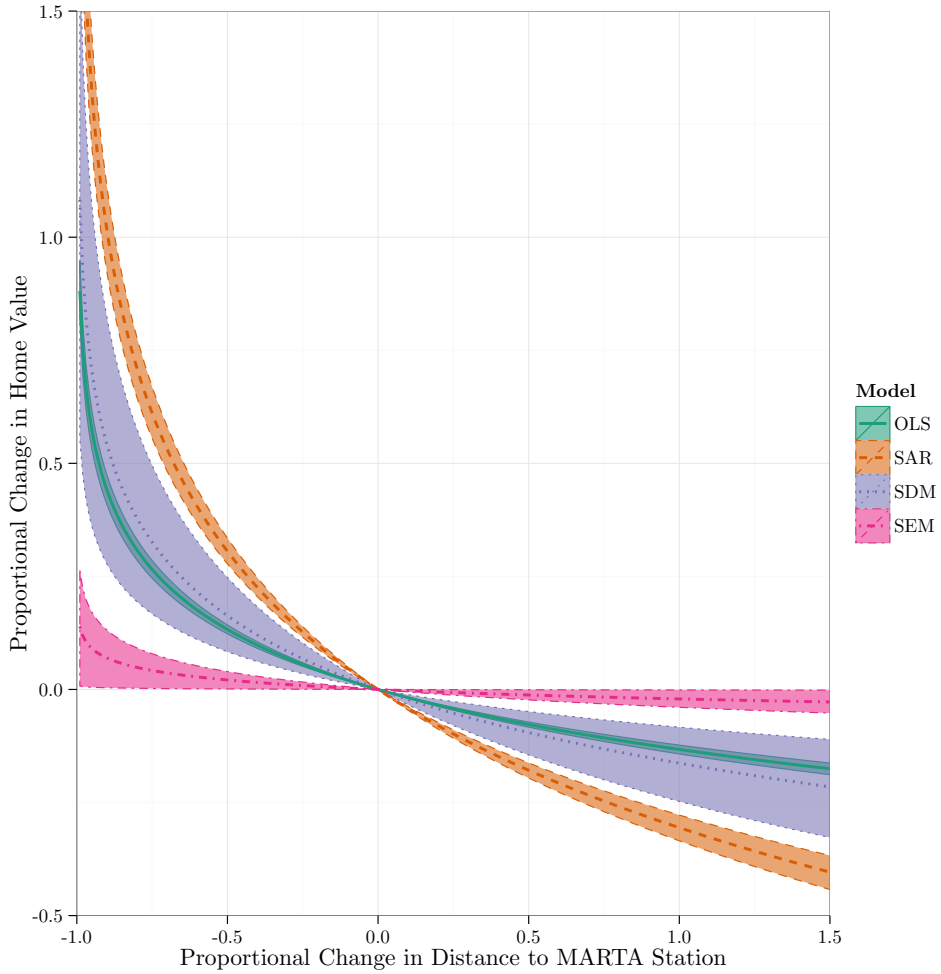

Figure 2: Estimated elasticity for transit proximity under different specifications, with confidence intervals.

in our case that the OLS and SEM models may have biased parameters, and that the OLS and SAR models may have unreliable confidence intervals.

These confidence intervals are crucial from a risk management perspective: a conservative land value capture forecast will use not the mean MWTP, but some lower percentile so that revenues are more likely to exceed the forecast. Consider a situation where a tax allocation district surrounding a transit line extension is forecast (using the mean OLS MWTP estimate of -0.192), to generate \$1 billion over the next 20 years.[3] A conservative estimate would be based on the 5^{th} percentile (-0.179) rather than the mean, and would generate \$930 million. But the 5^{th} percentile estimate from the SDM (-0.117) is only about 60% of the mean OLS estimate, implying that a truly conservative forecast would be in the neighborhood of \$600 million. This might have serious implications for the success of the project. That our 5^{th} estimate of the SDM effects is not meaningfully different from the mean OLS effect is fortunate, but should not be generally expected.

[3]This is in line with the forecasted revenue for the Atlanta Beltline tax allocation district.

Table 6: Comparison to related studies.

Statistic	Osland	Löchl and Axhausen	Ibeas et al.	This Study
$\ln(\mathcal{L}_{SDM})$	96.9	4192.6	-2.1	-272.7
$\ln(\mathcal{L}_{SEM})$	80.9	4118.0	-34.8	-338.1
$\ln(\mathcal{L}_{SAR})$	65.8		-75.0	-838.8
$\ln(\mathcal{L}_{OLS})$	52.0	3183.3	-111.9	-2006.7
$-2(\ln(\mathcal{L}_{SEM}) - \ln(\mathcal{L}_{SDM}))$	31.9*	149.2*	65.5*	130.9*
Classical	SEM**	SEM	Unknown†	SEM
General	SDM	SDM	SDM	SDM

* Reject null hypothesis with $p < 0.01$.

** Selected SDM after testing for common factors.

† Did not report LM statistics, but selected SEM.

5. Discussion

Autoregressive hedonic models that explicitly account for spatial dependence are not new to general econometrics and real estate science (Can and Megbolugbe, 1997; Dubin et al., 1999; Pace, 1997), and have been applied to estimate MWTP for transportation amenities. Haider and Miller (2000) showed spatial dependence was a significant issue in hedonic models of the Toronto market with respect to that city's transportation system. Armstrong and Rodríguez (2006) applied an SAR model to estimate the MWTP for access to commuter rail in the Boston suburbs, but did not comment on the effects of this dependence for their model estimates. Martínez and Viegas (2009) showed that MWTP estimates obtained with an SAR in Lisbon were similar to those obtained using OLS. In all of these cases, the authors used a single autogressive model, without an attempt to identify the preferred model.

Comprehensive analyses of spatial dependence and correlation together are increasingly common. Three recent studies in particular compare multiple autoregressive structures in a transportation or accessibility context. Osland (2010) selected the SDM after deciding that the LM tests were inconclusive, using the condition in line 21 of Algorithm 1. Löchl and Axhausen (2010) selected the SEM as the most appropriate model for a hedonic forecast in Zürich's UrbanSim land use model (Waddell et al., 2003), again applying the classical framework. In this case the LM tests were conclusive, with the RLM_λ test statistic about one hundred times greater than the RLM_ρ test statistic. Ibeas et al. (2012) similarly select the SEM model to estimate MWTP for transit accessibility in Santander; these authors do not report their LM test statistics, but they reject the SDM on account of some insignificant lagged parameters. According to the general selection framework, insignificant lagged parameters lead to the SAR model.

Would any of these authors have selected a different model with the general framework? The model likelihood values from each of these studies (including the present) are shown in Table 6. In all four cases, the common factors test rejects that the SDM and SEM have equivalent likelihoods, suggesting that the SDM would be the preferred model.

As mentioned in the Introduction and shown in Table 1, autocorrelation in the model residuals is a nuisance that affects the estimated standard errors of model parameters but not the parameter estimates themselves. Autocorrelation in the dependent variable, by contrast, is a substantive problem that will bias model parameters. Selecting the SEM when a SAR or SDM is the true model may result in biased parameters (reflected in Figure 2), whereas selecting an SDM when the SEM or the SAR is the true model merely sacrifices degrees of freedom to estimate unnecessary

parameters.

It is this last point that provides perhaps the greatest argument for the general framework. Standard null hypothesis significance testing is constructed to minimize the possibility of Type I error, or incorrectly rejecting a true null hypothesis. In the classical framework, the null hypothesis is that spatial effects are not present; the consequence of falsely rejecting this null hypothesis is an inefficient model. In the general framework, the consequences of falsely rejecting the null hypothesis of spatial effects are biased parameter estimates and/or invalid parameter significance tests. Further, the specification tests of (Larch and Walde, 2008) indicate that the general framework more frequently arrives at the true DGP. Beginning with the general model and reducing it when possible is the more conservative strategy.

6. Conclusion

Accurate estimates of MWTP for public transportation infrastructure are essential for regional transportation models and plans. It is therefore imperative that analysts select an econometric structure for their models that appropriately represents the complexity of the housing market that they seek to study. Spatial effects represent both a challenge and an opportunity for such models. If spatial dependence and correlation are not considered then estimates of MWTP may be unreliable. If spatial dependence and correlation are shown to exist, on the other hand, the analyst can use these effects to develop parsimonious and powerful models.

In this paper, we have shown that considering spatial dependence and correlation in the Atlanta housing market affects estimates of MWTP for proximity to MARTA. Specifically, the estimate MWTP *less certain*, implying that a land value capture strategy built on this model should consider a substantially higher margin of error in its forecasts.

The primary contribution of this paper is its comprehensive comparison of model selection frameworks, applying primarily theoretical advances to a problem commonly encountered by transportation practitioners. This comparison is particularly important in light of reduced funding for transportation infrastructure in the U.S. in particular, and the commensurate demand for improved planning tools and performance measures. The literature defining spatial effects and models to accommodate them is sufficiently mature that analysts should be aware of the risks stemming from spatial effects of various types, and be acquainted with tools to identify and address them. The re-orientation towards a general-to-specific framework will prevent analysts from incorrectly rejecting the conservative and general SDM in favor of a more efficient but potentially inappropriate specification.

References

Alonso, W., Jan. 1960. A theory of the urban land market. Papers in Regional Science 6 (1), 149–157.
URL http://doi.wiley.com/10.1111/j.1435-5597.1960.tb01710.x

Anselin, L., 1980. Estimation methods for spatial autoregressive structures. Regional Science Dissertation & Monograph Series, Program in Urban and Regional Studies, Cornell University (8).
URL http://www.cabdirect.org/abstracts/19801873236.html

Anselin, L., Sep. 1988a. Lagrange multiplier test diagnostics for spatial dependence and spatial heterogeneity. Geographical Analysis 20 (1), 1–17.
URL http://doi.wiley.com/10.1111/j.1538-4632.1988.tb00159.x

Anselin, L., 1988b. Spatial Econometrics: Methods and Models (Studies in Operational Regional Science). Kluwer, Dordrecht.

Anselin, L., Apr. 2003. Spatial externalities, spatial multipliers, and spatial econometrics. International Regional Science Review 26 (2), 153–166.
URL http://irx.sagepub.com/cgi/doi/10.1177/0160017602250972

Anselin, L., Bera, A. K., Florax, R. J. G. M., Yoon, M. J., 1996. Simple diagnostic tests for spatial dependence. Regional Science and Urban Economics 26, 77–104.

Armstrong, R. J., Rodríguez, D. A., Jan. 2006. An evaluation of the accessibility benefits of commuter rail in eastern Massachusetts using spatial hedonic price functions. Transportation 33 (1), 21–43.
URL http://www.springerlink.com/index/10.1007/s11116-005-0949-x

Atlanta Regional Commission, 2012. ARC GIS data and maps.
URL http://www.atlantaregional.com/info-center/gis-data-maps

Bivand, R., Jan. 2006. Implementing spatial data analysis software tools in R. Geographical Analysis 38 (1), 23–40.
URL http://doi.wiley.com/10.1111/j.0016-7363.2005.00672.x

Bowes, D. R., Ihlanfeldt, K. R., Jul. 2001. Identifying the impacts of rail transit stations on residential property values. Journal of Urban Economics 50 (1), 1–25.
URL http://linkinghub.elsevier.com/retrieve/pii/S0094119001922144

Brigham, E. F., 1965. The determinants of residential land values. Land Economics 41 (4), 325–334.

Brunsdon, C., Fotheringham, A. S., Charlton, M. E., 1999. Some notes on parametric significance tests for geographically weighted regression. Journal of Regional Science 39 (3), 497–524.

Burridge, P., 1981. Testing for a common factor in a spatial autoregression model. Environment and Planning A 13 (7), 795–800.
URL http://www.envplan.com/abstract.cgi?id=a130795

Can, A., Megbolugbe, I., 1997. Spatial dependence and house price index construction. Journal of Real Estate Finance and Economics 14, 203–222.

Chen, H., Ruffalo, A., Dueker, K. J., 1998. Measuring the impact of light rail systems on single-family home values: a hedonic approach with geographic information system application. Transportation Research Record 1617, 38–43.

Chernobai, E., Reibel, M., Carney, M., Oct. 2009. Nonlinear spatial and temporal effects of highway construction on house prices. The Journal of Real Estate Finance and Economics 42 (3), 348–370.

Cliff, A. D., Ord, J. K., 1970. Spatial autocorrelation: a review of existing and new measures with applications. Economic Geography 46, 269–292.

Debrezion, G., Pels, E., Rietveld, P., Jun. 2007. The impact of railway stations on residential and commercial property value: a meta-analysis. The Journal of Real Estate Finance and Economics 35 (2), 161–180.
URL http://link.springer.com/10.1007/s11146-007-9032-z

Dubin, R. A., Dec. 1998. Spatial autocorrelation: a primer. Journal of Housing Economics 7 (4), 304–327.
URL http://linkinghub.elsevier.com/retrieve/pii/S1051137798902364

Dubin, R. A., Pace, R. K., Thibodeau, T. G., 1999. Spatial autoregression techniques for real estate data. Journal of Real Estate Literature 7, 79–95.

Dubin, R. A., Sung, C.-H., Jun. 1987. Spatial variation in the price of housing: rent gradients in non-monocentric cities. Urban Studies 24 (3), 193–204.

Florax, R. J. G. M., Folmer, H., Rey, S. J., 2003. Specification searches in spatial econometrics: The relevance of Hendry's methodology. Regional Science and Urban Economics 33, 557–579.

Grass, R. G., 1992. The estimation of residential property values around transit station sites in Washington, D.C. Journal of Economics & Finance 16 (2), 139–146.

Haider, M., Miller, E. J., 2000. Effects of transportation infrastructure and location on residential real estate values: application of spatial autoregressive techniques. Transportation Research Record 1722, 1–8.

Hendry, D. F., 1979. Predictive failure and econometric modelling in macroeconomics: the transactions demand for money. Economic Modelling, 217–242.

Hess, D. B., Almeida, T. M., May 2007. Impact of proximity to light rail rapid transit on station-area property values in Buffalo, New York. Urban Studies 44 (5), 1041–1068.
URL http://usj.sagepub.com/cgi/doi/10.1080/00420980701256005

Iacono, M., Levinson, D., Dec. 2011. Location, regional accessibility, and price effects: evidence from home sales in Hennepin County, Minnesota. Transportation Research Record 2245, 87–94.
URL http://trb.metapress.com/openurl.asp?genre=article&id=doi:10.3141/2245-11

Ibeas, A., Cordera, R., Dell'Olio, L., Coppola, P., Dominguez, A., Sep. 2012. Modelling transport and real-estate values interactions in urban systems. Journal of Transport Geography 24, 370–382.
URL http://linkinghub.elsevier.com/retrieve/pii/S096669231200124X

Kim, J., Goldsmith, P., Jul. 2008. A spatial hedonic approach to assess the impact of swine production on residential

property values. Environmental and Resource Economics 42 (4), 509–534.
URL http://www.springerlink.com/index/10.1007/s10640-008-9221-0

Kressner, J. D., Garrow, L. A., Dec. 2012. Lifestyle segmentation variables as predictors of home-based trips for Atlanta, Georgia, airport. Transportation Research Record 2266, 20–30.
URL http://trb.metapress.com/openurl.asp?genre=article&id=doi:10.3141/2266-03

Larch, M., Walde, J., 2008. Lag or error? Detecting the nature of spatial correlation. In: Preisach, C., Burkhardt, H., Schmidt-Thieme, L., Decker, R. (Eds.), Data Analysis, Machine Learning and Applications. Springer, Berlin, Freiburg, pp. 301–308.

LeSage, J. P., Fischer, M. M., Nov. 2008. Spatial growth regressions: model specification, estimation, and interpretation. Spatial Economic Analysis 3 (3), 275–304.
URL http://www.tandfonline.com/doi/abs/10.1080/17421770802353758

LeSage, J. P., Pace, R. K., 2009. Introduction to Spatial Econometrics. Chapman and Hall/CRC.

Lewis-Workman, S., Brod, D., 1997. Measuring the neighborhood benefits of rail transit accessibility. Transportation Research Record 1576, 147–153.

Löchl, M., Axhausen, K. W., 2010. Modeling hedonic residential rents for land use and transport simulation while considering spatial effects. Journal of Transport and Land Use 3 (2), 39–63.

Martínez, L. M., Viegas, J. M., Dec. 2009. Effects of transportation accessibility on residential property values. Transportation Research Record 2115, 127–137.
URL http://trb.metapress.com/openurl.asp?genre=article&id=doi:10.3141/2115-16

Massell, B. F., Stewart, J. M., 1971. The determinants of residential property values. Institute for Public Policy Analysis, Stanford University, Discussion Paper (6).

McMillen, D. P., McDonald, J., 2004. Reaction of house prices to a new rapid transit line: Chicago's Midway Line, 1983-1999. Real Estate Economics 32 (3), 463–486.

Mikelbank, B. A., Dec. 2004. Spatial analysis of the relationship between housing values and investments in transportation infrastructure. The Annals of Regional Science 38 (4), 705–726.

Mur, J., Angulo, A. M., Nov. 2006. The spatial Durbin model and the common factor tests. Spatial Economic Analysis 1 (2), 207–226.
URL http://www.tandfonline.com/doi/abs/10.1080/17421770601009841

Nelson, A., 1992. Effects of elevated heavy-rail transit stations on house prices with respect to neighborhood income. Transportation Research Record 1359, 127–132.
URL http://trid.trb.org/view.aspx?id=371633

Osland, L., 2010. An application of spatial econometrics in relation to hedonic house price modeling. The Journal of Real Estate Research 32 (3), 289–320.

Pace, R. K., Jul. 1997. Performing large spatial regressions and autoregressions. Economics Letters 54 (3), 283–291.
URL http://linkinghub.elsevier.com/retrieve/pii/S0165176597000268

Páez, A., Scott, D. M., Volz, E., Oct. 2008. Weight matrices for social influence analysis: An investigation of measurement errors and their effect on model identification and estimation quality. Social Networks 30 (4), 309–317.
URL http://www.sciencedirect.com/science/article/pii/S0378873308000282

R Development Core Team, 2013. R: A Language and Environment for Statistical Computing.
URL http://www.r-project.org

Ridker, R. G., Henning, J. A., 1967. The determinants of residential property values with special reference to air pollution. The Review of Economics and Statistics 49 (2), 246–257.

Rosen, S., 1974. Hedonic prices and implicit markets: product differentiation in pure competition. The Journal of Political Economy 82 (1), 34–55.

Smith, J. J., Gihring, T. A., Jul. 2006. Financing transit systems through value capture. American Journal of Economics and Sociology 65 (3), 751–786.
URL http://doi.wiley.com/10.1111/j.1536-7150.2006.00474.x

Tax Policy Center, 2013. What are the sources of revenue for local governments?
URL http://www.taxpolicycenter.org/briefing-book/

Tobler, W. R., 1970. A Computer Movie Simulating Urban Growth in the Detroit Region. Economic Geography 46, 234–240.

U.S. Census Bureau, 2010. Wealth and Asset Ownership.
URL http://www.census.gov/people/wealth/

Waddell, P. A., Borning, A., Noth, M., Freier, N., Becke, M., Ulfarsson, G., Jan. 2003. Microsimulation of urban development and location choices: design and implentation of UrbanSim. Networks and Spatial Economics 3 (1),

43–67.
URL http://www.springerlink.com/content/jx04r0832r37x237

Appendix A. Sampling Bias

The sampling frame for the study data is the Georgia Motor Vehicles registration database. This raises the possibility that the sample is unrepresentative of households in the Atlanta region, as households owning multiple vehicles have a higher likelihood of being sampled, and households that do not own vehicles — or that only lease vehicles — are excluded. This is only a problem, however, if zero-vehicle households are common among home owners, as we excluded renters from our analysis. To examine the potential for unrepresentativeness in our sample, we compared our analysis data with the 2006-2011 5-year aggregated public use microdata sample (PUMS) file representing Fulton and DeKalb counties from the American Community Survey (ACS).

According to the PUMS data, 587 households of the 30,381 respondents in Fulton and DeKalb counties owned a home but did not own a vehicle, implying that we failed to sample approximately 3.02% of the relevant households. Our data contain a different set of variables than the ACS questionnaire, and we therefore cannot compare the datasets variable-to-variable. For the household income variable, however, we were able to run a Kolmogorov-Smirnov test comparing the distribution in our sample versus that in the ACS PUMS: we rejected that the two distributions were the same with a p-value of less than 1×10^{-16}. Figure 1 illustrates where our sample differs from the ACS microdata: we observe fewer households with incomes over \$250k, but more in the \$100k to \$175k range. The results of this comparison analysis suggest that our sample may not be perfectly representative of the Atlanta housing market, but not likely enough to seriously compromise our findings.

Appendix B. Spatial Weights

There are a number of ways to specify spatial weights matrices. The most common spatial simultaneous autoregressive models were originally developed for areal regions, and this has led to a number of specifications that are not primarily intuitive. There are three basic species of weights matrix:

1. Adjacency: Do Voronoi polygons around the observations touch? This can be extended to higher orders.
2. Nearest k observations, regardless of distance.
3. Observations within distance d, regardless of number.

The nearest observations and distance radius methods may both be weighted by distance to assign higher value to nearer observations. This creates five candidate schema, each with an array of inclusion possibilities (by allowing k or d to increase).

We examine the model likelihood and parameter stability of a spatial Durbin model using each of the five candidate schema with nine different inclusion rules. For the Voronoi polygons, we considered 1^{st} through 9^{th}-order adjacency. For the k -nearest neighbors method, we use $k =$ 2,5,10,15,20,35,50,75, and 100. For the radius method, we use nine equal divisions of the range $d = [0.5, 4.0]$ miles. We also consider an inverse distance weighting scheme for both the k nearest and d radius schema.

Figure 1 shows the log-likelihood of our SDM specification estimated for each of the candidate weights matrices. As shown in the figure, the Voronoi polygon method produces its maximum log-likelihood for first-order contiguity, and drops substantially as higher orders are considered. The k nearest neighbors method produces its maximum at 20 neighbors, a much higher level than

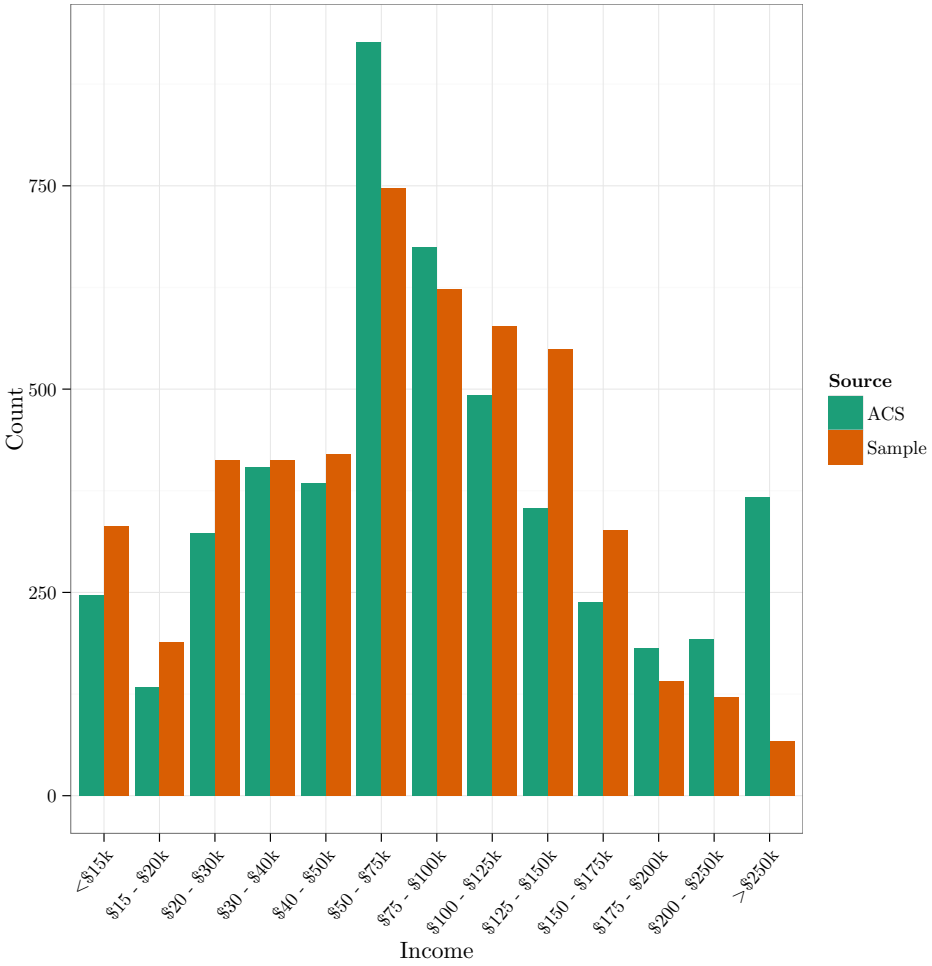

Figure 1: Distributions of incomes in the ACS and our estimation sample.

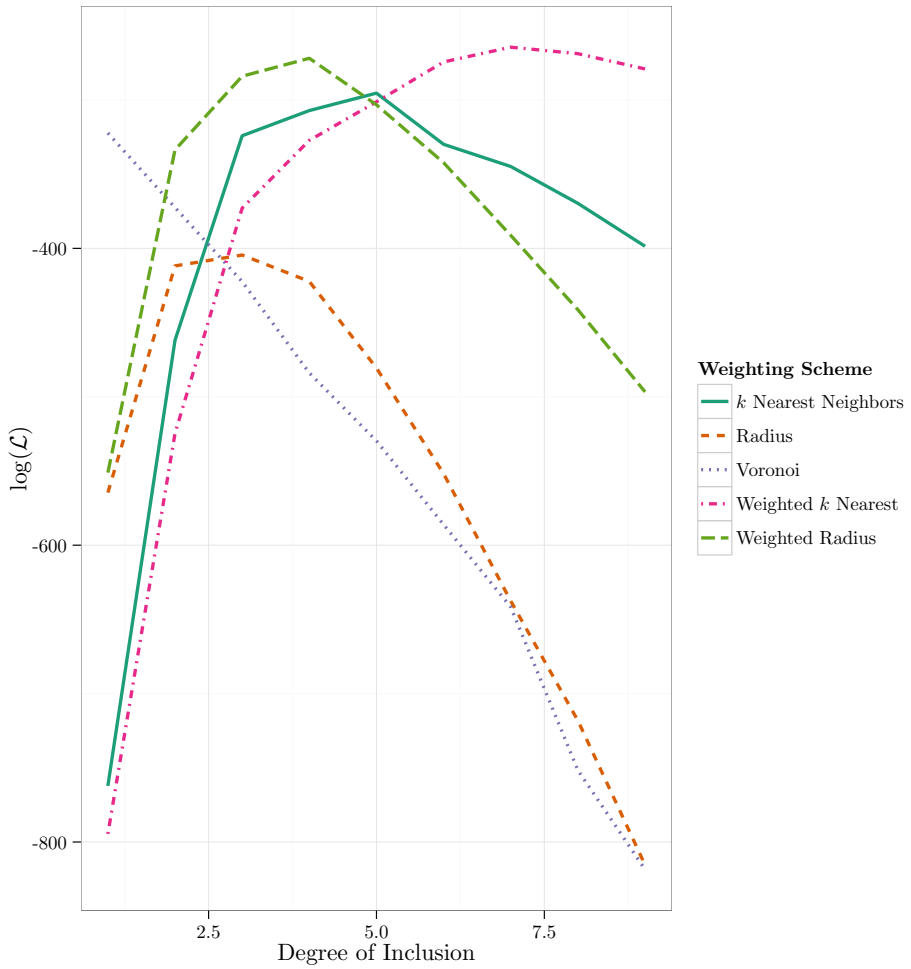

Figure 1: Log-likelihood under different weighting regimens.

used by either Löchl and Axhausen (2010) or Ibeas et al. (2012). Considering neighbors within a particular radius has its highest likelihood at 3.1. Weighting both the k-nearest and d-radius methods substantially improves the maximum achieved likelihood for both methods, with the optimum number of neighbors being 50 and the optimum radius now 1.81 miles.

LeSage and Pace (2009) assert that the particular weighting scheme should not have a serious influence on the estimated parameters. Our findings presented in Figure 2 provide some initial support for that claim, but also some disputations. The autocorrelation parameter ρ increases monotonically with expanding inclusion, with the notable exception of the Voronoi polygon method, which drops drastically above 7^{th}-order adjacency. The direct coefficient β and the indirect coefficient γ are usually opposites; for instance, the weighted k method has the most positive β but also the most negative γ, potentially muting its effect. We selected the weighted radius method because it has a high model likelihood and conservative coefficent estimates, falling as they do in the middle of the range defined by the candidate weighting scheme. Establishing which scheme best represents a particular housing market, and under what conditions, is an important opportunity for further research.

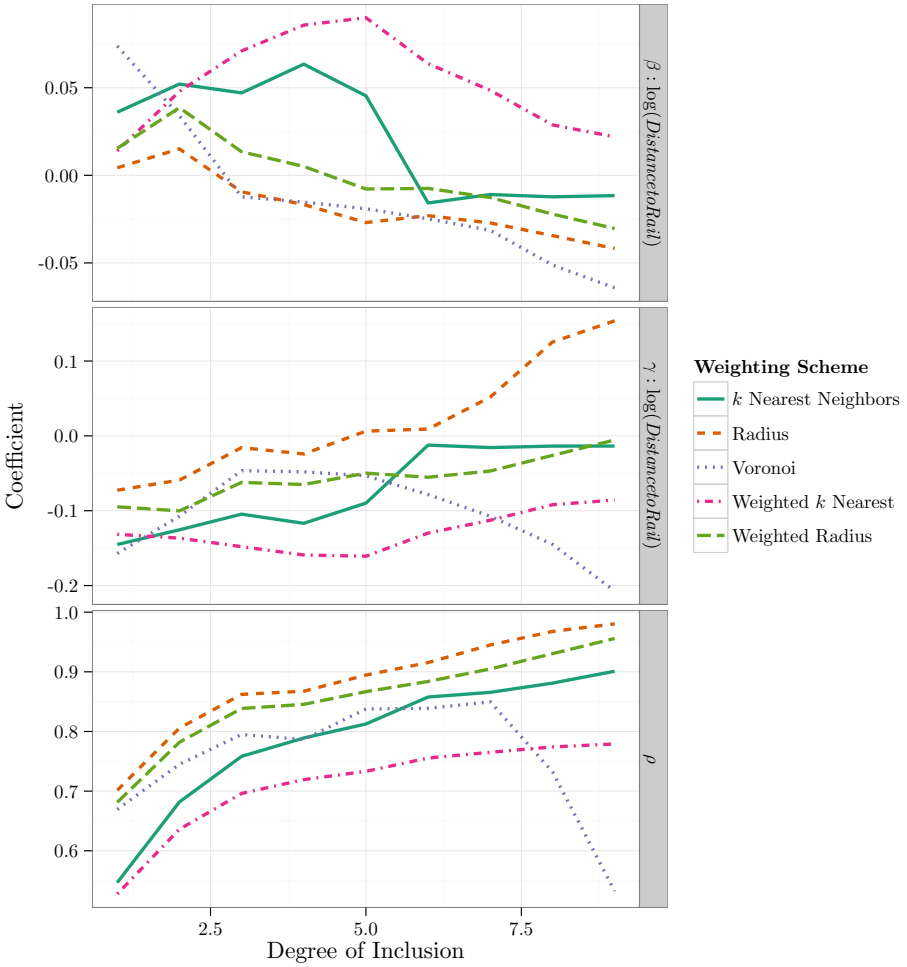

Figure 2: Autocorrelation and distance to rail parameters under differing weighting regimens.