Are Travel Demand Forecasting Models Biased because of Uncorrected Spatial Autocorrelation?

By

Frank Goetzke

RESEARCH PAPER 2003-10

Graduate Research Assistant
Department of Economics
College of Business and Economics
West Virginia University
Morgantown, WV 26506-6025
Tel: (937) 298-1427
Email: fgoetzke@yahoo.com

Paper presented at the North American Meeting of the
*Regional Science Association Internationa*l, Philadelphia, PA,
November 20-22, 2003

ABSTRACT: This paper discusses spatial autocorrelation in mode choice models, including what kind of bias it introduces and how to remedy the problem. The research shows that a spatially autocorrelated mode choice model, not uncommon because of, in terms of transit characteristics homogeneous neighborhoods, systematically overestimates transit trips from suburban transit-unfriendly areas and underestimates transit trips in the transit-friendly city center. Adding a spatial lag term into the model specification avoids the bias, however, it also changes sampling approaches, requires higher quality household forecast data and complicates forecasting.

1. INTRODUCTION

Pickrell's controversial study on new rail starts finds strong evidence that planners not only underestimated operation and capital cost, but also overestimated ridership for new federally funded rail system (Pickrell 1989). Kain (1992) supports these findings. Along the lines of the above authors, this paper discusses a potentially systematic bias in transit ridership forecasts, which stems from spatial autocorrelation in the mode choice model. Failure to account for spatial autocorrelation would lead to overestimates of ridership originating in nontransit-friendly, mostly suburban, neighborhoods and underestimates of ridership from transit-friendly urban areas. Since new federally funded rail projects, studied by Pickrell, typically serve the commuter market from the suburbs to the CBD, this bias could explain at least some of the inflated ridership estimates.

Why can a mode choice model be spatially autocorrelated? Like in time series data, where the dependent variable might be a function of its own time lag, the dependent variable in spatial data series might be a function of its own spatial lag as well. Spatial autocorrelation is often the case when neighborhoods are homogeneous. For instance, since transit ridership depends on both service attributes and accessibility, which are not expected to differ significantly between neighboring areas, it is easy to imagine that the knowledge of mode share in surrounding areas adds information for determining mode share in the study area. Therefore, spatial autocorrelation would be empirically supported if mode share changes gradually and not abruptly from one zone to another, and if zones with high transit mode share and low transit mode share are clustered together as it is the

case with the CBD and suburbs, respectively. Data is not spatially correlated, if the spatial distribution of mode share is a random event.

Few spatially autocorrelated discrete choice models have been estimated because of its computational complexity (Pinske et al. 1998). However, as will be shown later in this paper, some "tricks" can be used to circumvent the problems of estimating a full spatially autocorrelated model.

The next section shows theoretically how a spatially autocorrelated mode choice model would lead to a biased estimate. After that, a potential solution for estimating a spatially autocorrelated mode choice model is introduced in general terms. The following sections then discuss some technical issues of the spatially autocorrelated mode choice model, such as weight matrices, sampling and estimation. Ultimately, three tests to find evidence of spatial autocorrelation in mode choice data are shown and forecasting is discussed. The paper will conclude with some remarks regarding future empirical research needs.

2. WHAT IS THE PROBLEM?

The problem can be best described as an omitted variable bias. To see that, consider the following true model

$$v = x\,\beta + z\,\mu + \varepsilon \qquad (1)$$

where v is a latent variable, such as the utility from transit use, x is an $n \times 1$ vector of one mode choice determining socio-economic or trip characteristic, such as income, age, gender, auto availability, in-vehicle travel time, out-of-vehicle travel time, wait time, trip cost or distance to transit, n is the number of household observations, β is the regression coefficient, and z is an $n \times 1$ vector of an omitted, but relevant neighborhood variable, such as "transit-friendliness", which is difficult to measure empirically with μ the corresponding regression coefficient. The error term is ε with $E(\varepsilon) = 0$ and $Var(\varepsilon) = \sigma^2$.[1] The estimated model, which does not include the omitted variable z, looks as follows:

$$v = x\beta + \varepsilon \qquad (2)$$

The estimator b for the regression coefficient β is:

$$\begin{aligned} b &= (x'x)^{-1} x'v \\ &= (x'x)^{-1} x'(x\beta + z\mu + \varepsilon) \\ &= [(x'x)^{-1} x'x]\beta + [(x'x)^{-1} x'z]\mu + (x'x)^{-1} x'\varepsilon \\ &= \beta + [(x'x)^{-1} x'z]\mu + (x'x)^{-1} x'\varepsilon \end{aligned} \qquad (3)$$

And the expected value for the estimator b is:

$$\begin{aligned} E(b) &= E\{\beta + [(x'x)^{-1} x'z]\mu + (x'x)^{-1} x'\varepsilon\} \\ &= E(\beta) + E\{[(x'x)^{-1} x'z]\mu\} + E[(x'x)^{-1} x'\varepsilon] \\ &= \beta + [(x'x)^{-1} x'z]\mu \\ &= \beta + [Cov(x, z)/Var(x)]\mu \end{aligned} \qquad (4)$$

[1] To limit matrix algebra operations and make the argument more understandable, the model was chosen in such a way that there is only a vector x with one independent variable, besides the omitted variable, instead of a matrix X with several independent variables. The conclusion of the paper is also valid with more than one RHS variable (again, besides the omitted one), the math to show this theoretically, however, would become more complicated.

The estimator b is biased by the term $[Cov(x, z)/Var(x)]\ \mu$. Assuming that the latent variable v is disutility for using transit, z is chosen so that $\mu > 0$[2] and x is transformed in so that $Cov(x, z) > 0$[3], then the bias term $[Cov(x, z)/Var(x)]\ \mu > 0$ and the estimator b will be upwardly biased (see Equation 4).[4] As a result, an individual v_i takes on only the right value as long as $z_i = z^*$ for all observations $i = \{1, \ldots, n\}$ and with an average transit-friendliness $z^* = (\Sigma\ z_i)/n$. The bias of the estimator b adjusts the whole regression to improve the overall fit of the model, while omitting the term $(z\ \mu)$. If $z_i < z^*$, which would be the case in the transit-friendly central city, it can be seen that disutility v_i will be overestimated, causing the estimated mode choice for transit use to be lower than in reality. In suburban locations, on the other hand, where $z_i > z^*$, v_i will be too low, which biases the outcome to a higher number of estimated transit users compared to the true ridership.

Figure 1 graphically displays the relationship between v_i and z_i. It is obvious, that, if the estimated mode choice model has omitted a variable such as transit-(un)friendliness, then the estimated regression coefficient is consistently biased, and the model would overestimate transit use in the less transit-friendly areas, as it is the case in suburbs.

This is a far reaching discovery, since mode choice models likely depends on a factor such as "transit-(un)friendliness", which is extremely difficult and costly to

[2] In above case z would be increasing if the environment becomes more transit-unfriendly.
[3] If $Cov(x, z) > 0$ and z are chosen that $\mu < 0$, then $Cov(v, z) < 0$ and $Cov(x, v) < 0$. This means that x is transformed so that a smaller x_i increases v_i and, therefore, that all x_j are inverse measures of transit support such as income or travel cost/time.
[4] Similar arguments, with the same results, can be made for the other cases $\mu > 0$ and $Cov(x, z) < 0$, or $\mu < 0$ and $Cov(x, z) > 0$ and $Cov(x, z) < 0$, respectively.

quantify. Typically used approximations of transit-(un)friendliness, e.g. residential density and distance to transit are poor proxies and still leave an omitted variable bias.

Most major new public transit projects in the last 25 were light or heavy rail lines connecting the suburbs with the CBD, offering a commuting alternative for professionals living outside of the city, but working in the city center. A consistent bias resulting in overestimating transit riders in the suburbs could lead to significantly wrong forecasts that incorrectly encourage support for new rail construction.

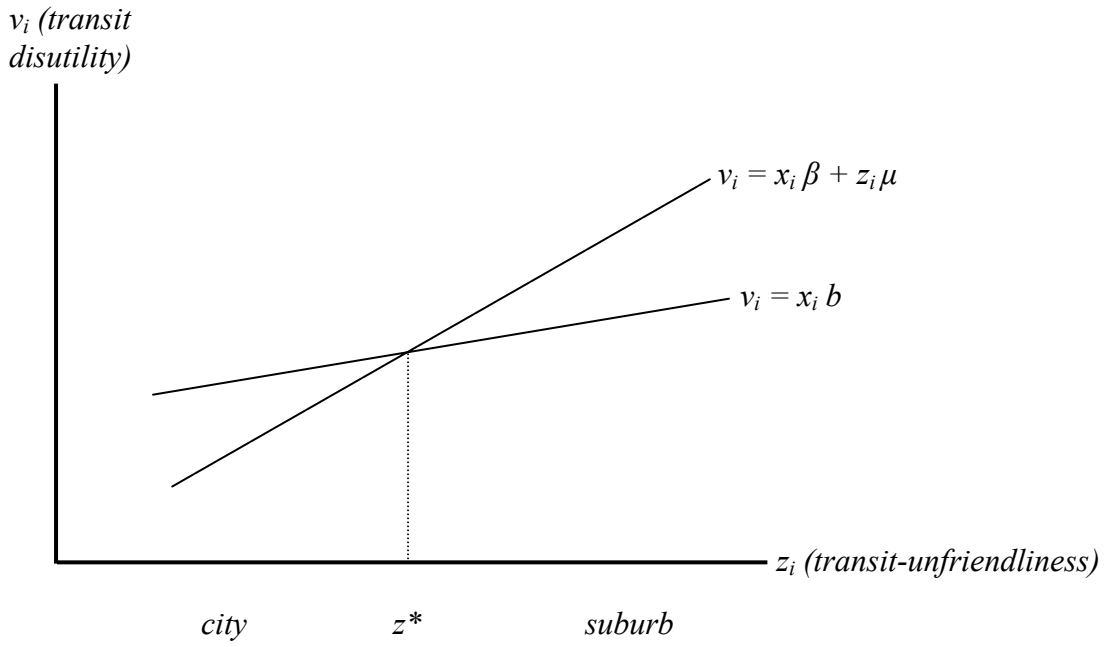

FIGURE 1: Relationship between transit disutility v_i and transit-(un)friendliness z_i of the household observation i for the true model including the variable z_i, as well as the estimated model omitting the variable z_i.

3. THE REMEDY

Again, let the true model be as shown in Equation 1: $v = x\,\beta + z\,\mu + \varepsilon$. This time, however, a spatial lag model is estimated:

$$v = \rho\, W v^* + x\,\beta + \varepsilon \qquad (5)$$

with v^* being the actual mode choice which is a function of the latent variable v, W a spatial weight matrix and ρ the regression coefficient for the spatial term. Then the estimator, r, of the regression coefficient, ρ, is the following expression:

$$\begin{aligned} r &= [(W v^*)'\,(W v^*)]^{-1}\,(W v^*)'\,(v - x\,b) \\ &= [(W v^*)'\,(W v^*)]^{-1}\,(W v^*)'\,(x\,\beta + z\,\mu + \varepsilon - x\,b) \end{aligned} \qquad (6)$$

If $(W v^*)$ is strongly correlated with the omitted variable z, which is not unlikely, then b is unbiased, since no variable is omitted anymore. This means that $E(b) = \beta$ and

$$\begin{aligned} E(r) &= E\{[(W v^*)'\,(W v^*)]^{-1}\,(W v^*)'\,(x\,\beta + z\,\mu + \varepsilon - x\,b)\} \\ &= E\{[(W v^*)'\,(W v^*)]^{-1}\,(W v^*)'\,z\,\mu + [(W v^*)'\,(W v^*)]^{-1}\,(W v^*)'\,\varepsilon\} \\ &= E[(z'\,z)^{-1}\,(z'\,z)\,\mu + (z'\,z)^{-1}\,z'\,\varepsilon] \\ &= E[\mu + (z'\,z)^{-1}\,z'\,\varepsilon] \\ &= E(\mu) + E[(z'\,z)^{-1}\,z'\,\varepsilon] \\ &= \mu + (z'\,z)^{-1}\,z'\,\varepsilon \end{aligned} \qquad (7)$$

Equation 7 is like the regular regression coefficient outcome, except, as it is the case in all spatial regression models, that the estimator r has a simultaneous equation bias, because the error term $[(z'\,z)^{-1}\,z'\,\varepsilon]$ is correlated with the endogenous variable v and its expected value is non-zero. Therefore, a OLS regression estimation would be inconsistent. However, this is a discrete choice model, based on either a logit or probit approach,

which cannot be estimated with simple OLS techniques. The method to estimate this kind of model is maximum likelihood, resulting in a consistent value for *r*.

The procedure above can be compared with the instrumental variable procedure. But the spatial lag term *(W v*)* is both an instrumental variable, as well as a "reversed" instrumental variable. *(W v*)* is an instrumental variable in the sense that it is an instrument for *z*, which is difficult to get data on. On the other hand, *(W v*)* is also correlated with the endogenous variable, which is in some way the opposite of an instrumental variable that would be used to avoid correlation with the LHS.

4. THE SPATIAL WEIGHT MATRIX

One important practical issue concerns how to design a spatial weight matrix *W* that has purely the purpose to define the spatial relation among households and is used to calculate the weighted average of the actual mode choice *v**. For travel demand forecasting models the whole study region is typically divided into hundreds or even thousands of travel analysis zones (TAZ), which are unique geographic areas, often aggregated census tracts with a similar number of households as well as compatible spatial attributes. Each household used for estimating mode choice belongs in one TAZ. However, a TAZ has more than one sampled household.

If, in fact, each household would be its own zone, then the weight matrix would be based on the traditional spatial rook or queen contiguity between the locations of the

households.[5] With more than one household in each TAZ, however, neighborhood relations between households have to be defined somewhat differently to what is tradition in spatial analysis.[6] The easiest way to construct a spatial weight matrix is to use all the households in the same TAZ as immediate neighbors, but none of the households in any other TAZ. This leads to another problem. Since the household data is sampled and, thus, does not represent the complete population, it makes sense, against the convention of spatial weight matrices, to make each household also its own neighbor. Consequently, the weight matrix has values on its main diagonal axis. As long all households belonging to the same TAZ are next to each other, the weight matrix W will consist of diagonal blocks with values, because households are only spatially correlated within the same zone. With increasing sample size in each zone, it does not make any difference whether to include the main diagonal values or not.

When $(W\ v^*)$ is calculated based on the above weight matrix definition, it can be found that it is just the zonal average of v^*. For instance, taking transit mode choice for v^*, $(W\ v^*)$ would be the percent of transit use in the TAZ. It is indeed conceivable that this average mode choice is strongly correlated with transit-(un)friendliness. Another way to justify the use of $(W\ v^*)$ is best expressed allegorically: Transit use is contagious. This can be thought of in two ways:

- A household is more likely to use transit if all its neighbors use transit, because the conditions of using transit in the area are good and the household will find out.

[5] Rook contiguity is defined as two neighbors sharing a common border. Queen contiguity adds a spatial corner relationship between two neighbors without a common border. An example for a queen contiguity would be the two states Arizona and Colorado.
[6] I want to emphasize that there is no right weight matrix, only better and worse ones.

- If people plan to use transit, they are more likely to move into an area where other people already use transit, because, again, the conditions of using transit are good.

Another, more complex spatial weight matrix design would be to include all the households of neighboring TAZs as well, following either a rook or queen relationship. The result would be an average of mode choice v^*, weighted by a factor based on the own TAZ and all of the neighboring TAZs. These weights could come from the number of households in each zone, the size of the TAZ, the length of the border, or they could be just a generic value, such as 50% for the own TAZ and the remaining 50% distributed equally between the neighboring zones. As seen above, the main diagonal of W will again contain values.

The last possible approach for a weight matrix is to not use TAZs at all, and instead to take distances between the sampled households. However, this requires more geographic information, such as x and y coordinates, which should not be difficult to get but might violate the privacy rights of the sampled household. As discussed in the next section, it will also put additional stipulations on sampling and is not very practical.

5. SPATIAL SAMPLING

Without any spatial structure, a random sample should be a representative of the population. If it is known that the population is spatially autocorrelated, as it is expected with virtually any spatial data sets, then the sample should reflect this fact. This requires not only that the mean and standard deviations of each characteristic in the sample for the

whole study area is equal to the mean and standard deviation of each characteristic in the complete population, but also, that this is true for all spatial subareas (TAZs). Therefore, sampling for household surveys needs to be done on the "lowest" spatial level, e.g. zones. In practice the difference between standard and spatial sampling is that the survey participants are not drawn from the population of all households, but instead from the spatial subset of this population, such as zones.

Standard deviations for household characteristics should be smaller at the zone level compared to the whole study area. Therefore, the sample size for each zone can be less than it would be for the study area. And it is not expected that the total sample size will increase significantly, but rather that the observations will have a more even spatial distribution.

In order to make a conceptually different distance based weight matrix compared to a spatial contiguity weight matrix, it is important that the sampled households are approximately equally spaced and that they are not clustered. The easiest sampling approach would be to lay a grid above the study area and draw from each grid square one or any other fixed number of households. This is similar to the stratified systematic unaligned point sample approach (McGrew et al. 1993).

6. THE ESTIMATION

A discrete choice model with a spatial lag term of the general form $y = \rho W y + x \beta + \varepsilon$ is difficult to estimate, because the latent variable y is recursive and, at the same time

cannot be directly observed. However, the model in Equation 5 is slightly different in the sense that, instead of the latent variable v, the observed LHS variable v^* is included on the RHS of the equation. This minor change makes a major difference for running the regression, because the model becomes what is in the literature called a conditional spatial discrete choice model. But, while estimating a conditional spatial discrete choice model is simpler to estimate, the trade-off is, that forecasting becomes more cumbersome.

A conditional discrete-choice model is best estimated with the regular logit or probit approach. The only difference is one additional term on the RHS, which is $\rho (W v^*)$. Therefore it is preferable to first compute $(W v^*)$ and then to run the regression to derive an estimator for ρ together with β. This can be done using any econometrics software capable of logit or probit model estimation.

7. THREE TESTS FOR SPATIAL AUTOCORRELATION

The goal is to account for all spatial autocorrelation with relevant independent variables, such as personal, household, neighborhood and/or commute characteristics. However, this cannot always be achieved. Furthermore, some relevant variables cannot be easily observed, such as transit-(un)friendliness. Therefore, the regression model might remain spatially correlated. How can it be known if the model is spatially correlated, or if all relevant variables are included? Three tests are discussed here to answer this question.

The first more general test is for a discrete-choice model modified version of the traditional Moran's I (Kelejian et al. 2001). This test is a spatial version of the Durbin-

Watson test and detects spatial autocorrelation by analyzing the residuals. The test statistics is normally distributed and looks like:

$$I = Q/\sigma_q \sim N(0, 1) \qquad (8)$$

Q is the weighted cross-product of the residuals $Q = (e' W e)$, and σ_q represents a normalization factor.

Alternatively, a different Moran's I can be used (Pinske et al. 1998). However, this time the statistics now follows the chi-square distribution:

$$LM = (u' W u)^2/T \sim Chi\text{-}square(1) \qquad (9)$$

In this version of Moran's I, the residuals u are standardized.

However, just like the Durbin-Watson test, both Moran's I statistics are unspecific and can therefore not to be trusted too much if a reasonable alternative exists.

The last, probably easier and more precise approach is to use an LR test for an omitted variable (Greene 1997). The principle behind the test is to compare the fit of the restricted model without the spatial lag variable (H_0) to the unrestricted model with the spatial autocorrelation term (H_A):

$$H_0: \quad v = x\beta + \varepsilon$$
$$H_A: \quad v = \rho W v^* + x\beta + \varepsilon \qquad (10)$$

Then the likelihood ratio of the restricted log-likelihood function (L_R) to the unrestricted function (L) follows a chi-square distribution with one degree of freedom for one restriction:

$$LR = -2 (\ln L_R - \ln L_U) \sim \text{Chi-square (1)} \qquad (11)$$

In summary, there are three different methods to test for existence of spatial autocorrelation in the model. Two modified Moran's I tests use the residuals of the model to detect spatial autocorrelation. The LR test compares the fit of the model with the spatial lag variable to the model without it, and finds if the non-inclusion of the spatial autocorrelation term is equivalent to omitting a relevant variable.

8. FORECASTING

Unfortunately, the conditional discrete choice model does not have a simple algebraic solution, as one finds in the unconditional model:

$$y = \rho W y + x \beta$$
$$y - \rho W y = x \beta$$
$$(I - \rho W) y = x \beta$$
$$y = (I - \rho W)^{-1} x \beta$$
$$y = (I + \rho W + \rho^2 W^2 + ...) x \beta$$
$$y = x \beta + \rho W x \beta + \rho^2 W^2 x \beta + ... \qquad (12)$$

The basic structure, however, is the same for both models. Disutility v will be some function of spatially lagged household, neighborhood and trip characteristics, summarized in x, as well as the error term:

$$v = f(\rho, W, \beta, x, \varepsilon) \qquad (13)$$

The solution for v and the forecast for actual mode choice v^* can only be obtained by simulation. The first step is to compute $v^1 = x\ b$ without the spatial lag term. Then the actual mode choice decision $v^1 * = g(v^1)$ can be derived depending on v^1. After that, v^1* will be used to calculate the second round of v and v^*, which is $v^2 = r\ W\ v^1 * + x\ b$ and $v^2 * = g(v_2)$. These above steps will be repeated until $v = v^{m-1} = v^m$, and final mode choice is then $v^* = g(v)$.

Equation 12 shows the spatial impact of the mode choice outcome will depend on the household, neighborhood and trip characteristics of all spatially lagged households. This makes sense, because a neighborhood's household composition adjusts after its transportation system has changed. At the same time, since household and trip characteristics have additional weight, accurate household forecasts may require the model to have the transportation system an endogenous variable. A "quick and dirty" shortcut would be to transfer the mode share value from similar neighborhoods to the neighborhood where the transportation system is changed, since, as seen above, $(\rho\ W\ v^*)$ represents the average percentage of people using transit.

9. CONCLUSION

This paper presented strong theoretical evidence that mode choice models can suffer from spatial correlation. This spatial autocorrelation will bias the model forecast systematically by overestimating transit ridership in transit-unfriendly neighborhoods, such as the suburbs, and underestimate transit ridership in transit-friendly central cities.

This paper also discussed the design of a spatial weight matrix and issues of spatial sampling, it showed three tests for how to detect the presence of spatial autocorrelation in discrete choice models, and it introduced an approach to account for spatial autocorrelation in mode choice models so that transit ridership can be forecast more accurately.

The final proof that mode choice model is spatially autocorrelated, however, must be left to empirical research. Using the 1990 PUMS data set for the New York City metropolitan area, which covers the states of Connecticut, New Jersey and New York and consists of 58 zones, the author currently analyses if work trip mode choice decisions are spatially autocorrelated.

It would be up to the transportation planning and travel demand forecasting community to incorporate the findings of this paper into their work in order to improve the quality of future transportation studies, which is crucial for informed and good decision making.

REFERENCES:

Greene, W. H., 1997. *Econometric Analysis.* Prentice Hall: Upper Saddle River, NJ.

Kain, J. F., 1992. "The Use of Strawmen in the Economic Evaluation of Rail Transport Projects," *American Economic Review 82*.

Kelejian, H. and I. Prucha, 2001. "On the Asymptotic Distribution of the Moran I Test Statistic with Applications," *Journal of Econometrics 104*.

McGrew Jr., J. C. and C. B. Montoe, 1993. *An Introduction to Statistical Problem Solving in Geography.* Wm. C. Brown Publisher: Dubuque, IA.

Pickrell, D. H., 1989. "Urban Rail Transit Projects: Forecasts vs. Actual Ridership and Costs," U.S. Department of Transportation, Transportation System Center, Cambridge, MA.

Pinske, J. and M. Slade, 1998. "Contracting in Space: An application of Spatial Statistics to Discrete-choice Models," *Journal of Econometrics 85*.