Applying Structural Equation Modeling to Public Transit Supply and Demand

John C. Handley
Palo Alto Research Center, Inc., 800 Phillips Road, MS 128-27E, Webster, NY 14580

Abstract
The relationship between scheduled capacity and concomitant demand in transit is of fundamental interest to transit planners. The expectation is that offered capacity in terms of more frequent service and more bus stops is consumed by customers. Thus relationship has been modeled at an aggregate level for whole transit agencies across regions but not at the operational level. This work studies the statistical relationships among operational capacity, demand and revenue using a year's worth of day-by-day trip-level data over 29 routes of a public transit system. Structural equation modeling is used to account for groups of collinear observational variables to identify and link three latent constructs, revenue capacity, demand, and productivity. We find that the endogenous variables revenue capacity and demand are positively, but weakly linked, showing that increased scheduled capacity is used. Yet there are variations route by route and those variations have no obvious relationship to route function (cross-town inner city versus suburban commuting).

1. Introduction

The relationship between public transit service capacity and ridership has been oft studied and suffers from no lack of opinion. Taylor et al. show, at the aggregated level of urban areas in the United States, service frequency and fares have a small, but significant effect on ridership; factors involving geography, population characteristics and economic factors dominate (1). Geographic factors include degree of urbanization, population density and region within the US. Population characteristics include political party composition, proportion of immigrants, etc. Economic factors cover personal wealth. Transit service supply is represented by vehicle services hours (see below). Their analysis used data from the National Transit Database (NTD) plus the US Census Bureau for the year 2000 for 265 urbanized areas. Each datum was an area with aggregated ridership statistics. From this perspective, one obtains variation at the area level. Yet even in the presence of external factors, some significant influence of per capita transit usage is reserved by fares and service frequency (26% of variation). One key aspect of this study is the segregation of external factors (geography, income, etc.) and those under control of policy makers (e.g, fares).

Thompson and Brown show a strong significant relationship between service availability (coverage, frequency) in 82 metropolitan service areas (MSA) and per capita transit trips (2). Their unit of observations were larger than Taylor et al. and were segmented into small (<500,000 residents), medium (500,000-1,000,000 residents), and large (>1,000,000 residents) MSAs. Their data comprised NTD and US Census Bureau data from 1990 and 2000 and the ridership variable of interest was the percent change in passenger miles per capita. Their hypothesis was that an increase in service coverage (a decrease in percentage change in ratio of population to service miles) would positively influence ridership (that is, adding more capacity relative to the population increases ridership). This effect was significant only for all data and all MSA segments. The same was found to be true for the ratio of vehicle miles to route miles. That is, adding more vehicles increased ridership. Interestingly, they also found a positive relationship for connectivity, represented as the proportion of routes not serving the central business district (CBD).

Kohn studied 85 Canadian transit systems and constructed a linear regression model that showed increased fares reduced ridership while increased vehicle hours (capacity) was associated with increased ridership (3). It remains unclear which is cause and which is effect; that is, whether increased capacity drives ridership or the capacity is built to handle increasing demand.

Syed and Khan analyzed a 1995 attitudinal survey regarding rapid bus transit in Ottawa, Canada using factor analysis and logistic regression (4). The extracted factors are used to predict a binary response of ride/no ride. They show that, in 1995, availability of bus information was the most important factor.

Taylor and Fink provide a thorough review of what is known to date (5). In their conclusion, they point out shortcomings of causal analyses, including aggregation, collinearity and latent effects. The work presented here attempts to address some of these albeit using data from a particular transit agency.

This study differs from previous work in that it focuses on the operational capacity of daily trips and demand. It addresses to what extent scheduled capacity is actually used. As in previous studies, we aim to uncover relationships among latent variables.

Structural equation modeling (SEM) is a common technique used to understanding travel behavior. Githui et al. use to identify latent factors in transit survey data from the city of Nairobi, Kenya. From 25 variables, they use exploratory factor analysis (described below) to infer three exogenous variables, Service Quality, Commuter Safety and Travel Cost and one endogenous variable, Commuter Satisfaction. They uncovered a linear relationship among the latent variables that indicated commuter satisfaction is positively influenced primarily by service quality, secondarily by safety and negatively by travel cost (6).

De Abrene e Silva et al. apply SEM to 2003 Origination-Destination travel survey data collected in the Greater Montréal Area (7). A SEM related land use and travel behavior and showed intuitive results (e.g., car travel frequency positively influenced by commuting distance, transit and non-motorized vehicles were negatively related showing competing mode choices).

Kim built a SEM to identify factors that influence mobility of the elderly near Seattle, Washington (8). The constructed model includes latent factors for mobility and urban form. There are numerous relationships between observed variables like ethnicity and education that are related to mobility, but no statistical significant relationship was found between latent factors.

These studies all use the Linear Structural Relations (LISREL) model, which has an accompanying software package with fitting and diagnostics functionality (9, 10). Here, a Bayesian approach is adopted, one which has more modeling flexibility and is quite easy to use in practice (11).

2. Data

We study data from a Computer-Aided Dispatch/Automatic Vehicle Location (CAD/AVL) system which system maintains trip-level performance data for reporting to the NTD. A trip is a route that is run according to a schedule. For example, Route 10 could be run every weekday starting at 7 am. Each time it is run, it is constitutes a separate trip. It would typically take an hour or so to run, use a single bus and cover several miles. It is considered a basic unit of revenue production (and schedule).

Table 1 gives the descriptions of the variables of interest (12). Results will be presented for all routes, but first, the analysis is illustrated for a single route, Route 10. Route 10 is a major east-west transit corridor and is one of the region's busiest. Summary statistics are shown in Table 2. The set of observations come from the CAD/AVL database and comprise trip measurements over 394 days (29 Aug 2010 to 27 Sept 2011) from a medium-sized city in the Northeast of the United States.

Table 1. Variables of interest with their definitions.

Variable	**Description**
num_stops_scheduled	number of bus stops assigned to this trip in the schedule data
vehicle_rev_miles	total amount of revenue miles or kilometers this bus traveled during this trip
vehicle_rev_hours	total amount of revenue hours it took the bus to run this trip
total_ons	load balanced number of passengers that boarded through all doors at all bus stops on this trip
total_offs	the load balanced number of passengers that alighted through all doors at all bus stops on this trip
max_load	load balanced maximum number of passengers that were on the vehicle at a given time during the trip
passenger_miles	total of all the distances in miles or kilometers traveled by each of the passengers on this trip.
passenger_hours	total for all the passengers of the amount of time (measured from door open to door open) each traveled on this trip. 'passenger_hours' includes dwell time at the layover prior to the start of the trip, assuming there is boarding activity.

The data were cleaned using following protocol. Cases where the vehicle revenue miles or vehicle revenue hours zero were removed and well as those with training drivers. Cases were removed where the vehicle hours were more than three hours and cases where the vehicle miles are less than two as these were deemed anomalous. Finally, logs were taken of passenger miles and passenger hours. The data were subsetted by route in order to speed up estimation (which can take several hours) and to compare parameter estimates by route.

Table 2. Summary statistics for Route 10 data, n = 23,953.

Variable	Average	Sample Std dev
num_stops_scheduled	104.810	36.466
vehicle_rev_miles	10.899	4.283
vehicle_rev_hours	1.033	0.390
total_ons	48.972	27.367
total_offs	47.934	26.147
max_load	27.172	15.088
log_passenger_miles	4.571	1.122
log passenger_hours	2.419	0.868

3. The Latent Model

The first step before constructing a structural equation model is exploratory factor analysis. Typically, the data are normalized by subtracting the average and dividing by the estimated standard deviation. Exploratory factor analysis is commonly used to group together observational variables with common latent factors for dimensionality reduction. Consider the following model,

$$Y = \Lambda\xi + \varepsilon \qquad (1)$$

where Y is a $p\times 1$ vector of (centered and scaled) measurements (in our case, $p = 8$), Λ is a $p\times q$ matrix of loadings which is to be estimated from the data, ξ is a $q\times 1$ vector of latent factors (also to be estimated), and ε is a $p\times 1$ random vector of errors independent of ξ, $\varepsilon \sim MVN(0,\Psi_{\varepsilon})$, $\Psi_{\varepsilon} = diag(\psi_{\varepsilon 1},\ldots$. Typically, $q < p$. The model in eq. 1 is underdetermined from data, so some additional criterion is used. We can write,

$$Y \sim MVN(0,\Sigma) \qquad (2)$$

where $\Sigma = \Lambda\Lambda^{T} + \Psi_{\varepsilon}$, $\mathrm{cov}(Y,\xi) = \Lambda$.

Given a data set, one can match the sample covariance $\hat{\Sigma}$ with $\Sigma = \Lambda\Lambda^{T} + \Psi_{\varepsilon}$. In practice, the estimate of Λ is rotated to increase the dispersion of its columns, which increases the ability to interpret the factors (this is the varimax approach).

Table 3. Estimated loadings and latent factors. Significant factor loadings are in bold, showing three distinct factors.

Variable	Latent factor 1	Latent factor 2	Latent factor 3
num_stops_scheduled	**0.856**	0.128	0.170
vehicle_rev_miles	**0.944**	0.170	0.158
vehicle_rev_hours	**0.776**	0.258	0.264
total_ons	0.250	**0.832**	0.491
total_offs	0.262	**0.881**	0.308
max_load	0.159	**0.674**	0.641
passenger_miles	0.311	0.374	**0.863**
passenger_hours	0.261	0.405	**0.799**

In Table 3, we identify three factors. Usually one considers loadings with absolute values over 0.5 to be significant. Our three latent factors are interpreted as revenue capacity, demand, and productivity (hence the labels in Table 1). Max load could have gone into either demand or productivity, but it has a greater loading in demand, so it is assigned there. (Including four factors produces an uninterpretable fourth factor with low loadings.)

Based on these estimated factors, we construct a plausible structural equation model to link these latent factors. Revenue capacity represents the ability of the system to capture revenue--bus stops where passengers can be picked up and the distance and hours buses are scheduled to run and accept fares. Demand is the realized demand, paying customers getting on the bus. This is represented by boardings, alightings and the maximum number of passengers on a bus on any given trip. Productivity is a combination of revenue capacity and demand. It is production in the sense of capacity that is generating revenue. It is measured by passengers miles and hours, essentially a tally of paying customers captured and transported; it measures how much of the service is actually used. Our hypothesized relationship is that productivity is a linear combination of revenue capacity and demand. We further hypothesize that demand is a function of capacity; that is capacity is positively correlated with demand. Transit planners would expect believe this to be true: increased capacity offered through scheduling is consumed.

Below are the observational equation (eq. 3) and structural equation (eq. 4).

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ y_8 \end{bmatrix} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \\ \beta_6 \\ \beta_7 \\ \beta_8 \end{bmatrix} + \begin{bmatrix} \lambda_1 & 0 & 0 \\ \lambda_2 & 0 & 0 \\ \lambda_3 & 0 & 0 \\ 0 & \lambda_4 & 0 \\ 0 & \lambda_5 & 0 \\ 0 & \lambda_6 & 0 \\ 0 & 0 & \lambda_7 \\ 0 & 0 & \lambda_8 \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \\ \eta \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \\ \varepsilon_8 \end{bmatrix} \qquad (3)$$

$$\varepsilon_i \sim N(0, \psi_{\varepsilon i}),\ i = 1, \ldots$$

$$\xi_2 = \alpha \xi_1 + \delta_{\xi_2}$$
$$\eta = \gamma_1 \xi_1 + \gamma_2 \xi_2 + \delta_\eta \qquad (4)$$

$$\xi_1 \sim N(0, \psi_{\xi_1})$$

$$\delta_{\xi_2} \sim N(0, \psi_{\delta_{\xi_2}})$$

$$\delta_\eta \sim N(0, \psi_{\delta\eta})$$

δ_η and δ_{ξ_1} δ are independent.

where ξ_1, ξ_2, and η correspond to the three latent factors revenue capacity, demand and productivity.

In this formulation, all latent variables have zero mean and all the observational variables maintain their original scales. This differs from some formulation that set equal to one a single lambda for each latent variable.

4. Results

Parameters were estimated using Just Another Gibbs Sampler (JAGS) called from ‘R’ (13). Code was adapted from examples associated with Lee (11). Markov Chain Monte Carlo (MCMC) convergence processes were monitored and found to be satisfactory. Residuals were also computed and found not to indicate structural issues. The model is deemed to adequately fit the data. Figure 1 shows the model with parameter estimates.

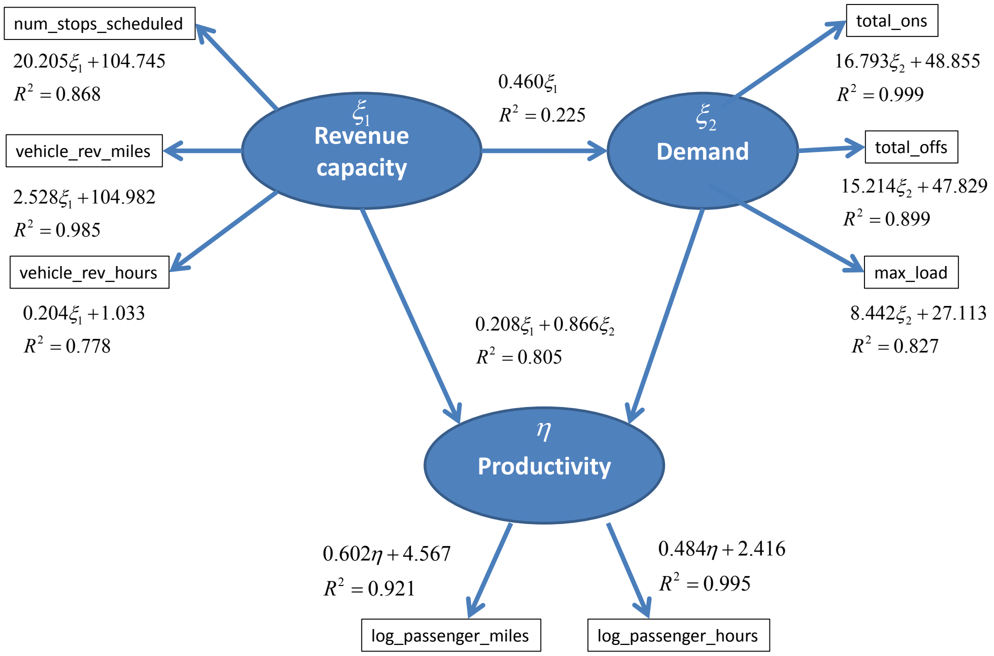

Figure 1. Structural equation model estimated for route 10, comprising 23,953 records over 13 months.

As specified by the model, the latent variables have estimated means of zero (Table 4). As expected, the estimated offsets β_i are close to the samples means in Table 2. This gives us some confidence in the structural part of the model.

Table 4. Posterior distributional properties of latent variable means.

Variable	Mean	SD	2.5%	Median	97.5%
ξ_1 (demand)	-0.007	0.006	-0.018	-0.008	0.004
ξ_2 (revenue capacity)	0.003	0.010	-0.018	0.003	0.022
η (productivity)	0.007	0.008	-0.008	0.007	0.022

As expected, observational variables are tightly related to their corresponding latent variables with R-squared values ranging from 0.788 to 0.999.

Demand is positively associated with revenue capacity ($\hat{\alpha} = 0.460$) and explains 22.5% of its variance. It is unclear whether the revenue capacity is built to respond to demand or vice versa. Nevertheless, the relationship is in the expected, positive direction. The model does not consider economic or demographic effects which one would expect to be greater influences on demand. Yet we do see a statistically significant relationship between what is deployed and what is used.

The second structural equation shows greater influence of demand than capacity on productivity, i.e., captured revenue ($\hat{\gamma}_1 = 0.208$, $\hat{\gamma}_2 = 0.866$). Again, it is expected that productivity, by its very definition of infrastructure use would be highly dependent on counts of boarding and alighting passengers. It is also evident that revenue capacity has a lesser, but significant role.

All routes were processed by subsetting the data. Figure 2 shows the estimated coefficients of determination (R-squared) for all routes along with 95% credible intervals.

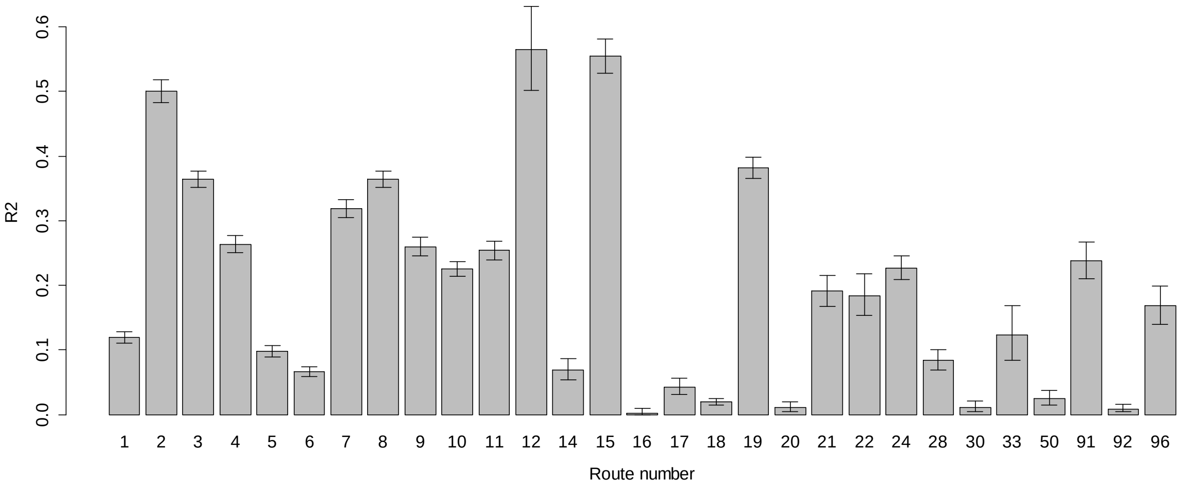

Figure 2. R-squared values with 95% credible intervals for $\xi_2 = \alpha\xi_1 + \delta_{\xi_2}$ for each route.

Routes 1-19 & 50 are urban and cross the central business district. Routes 20-33 are suburban and 91-96 are exurban and primarily serving commuters in adjacent counties with jobs downtown. There is no apparent relationship between route type and the degree to which revenue capacity influences demand.

Figure 3 shows the R-squared values for productivity as a linear function of revenue capacity and demand. Owing to the definition of productivity as a function of paying customers, these values are as expected, high for all routes.

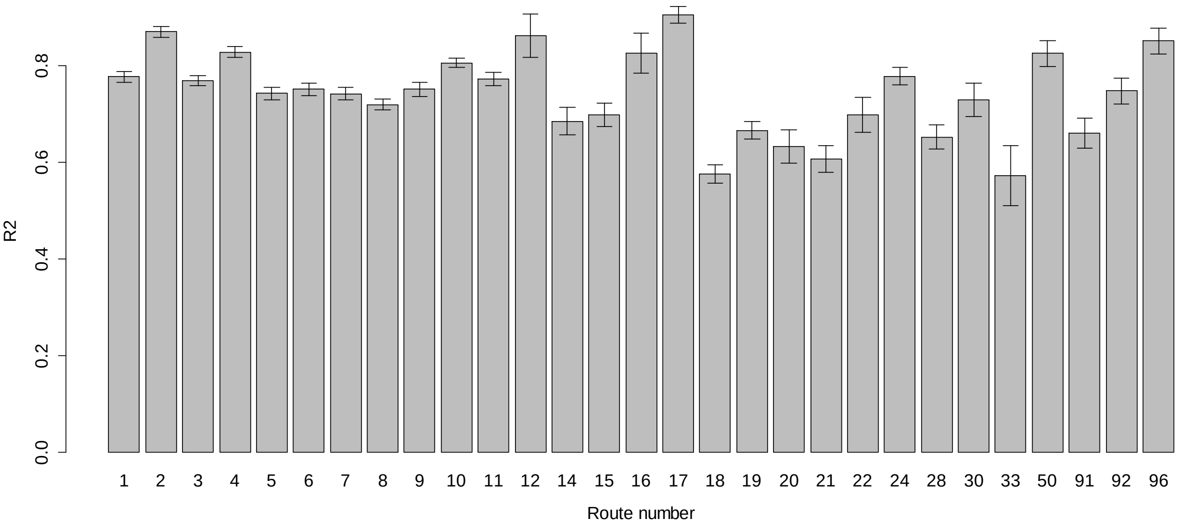

Figure 3. R-squared values with 95% credible intervals for $\eta = \gamma_1\xi_1 + \gamma_2\xi_2 + \delta_\eta$ for each route.

5. Conclusion

The link between transit capacity and usage has been studied at the aggregate level, to access policy and planning. The study presented here investigates the lower level relationship between that capacity which is scheduled and that which is used. We find an overall positive relationship, scheduled capacity is used day to day, but that relationship varies across routes and the variation bears no apparent relationship to route function. Structural equation modeling was used to elicit this relationship. A Bayesian approach afforded greater modeling flexibility and estimation of key statistical relationships.

References

1. Taylor, B. D., D. Miller, H. Iseki, and C. Fink. Nature and/or nurture? Analyzing the determinants of transit ridership across US urbanized areas. Transportation Research Part A: Policy and Practice, Vol. 43, No. 1, pp. 60-77, 2009.
2. Thompson, G. L. and J. R. Brown. Explaining variation in transit ridership in U.S. metropolitan areas between 1990 and 2000. Transportation Research Record: Journal of the Transportation Research Board, No. 1986, pp. 172-181, 2006.
3. Kohn, H. Factors affecting urban transit ridership. Proceedings, Bridging the Gaps Conference, Canadian Transportation Research Forum, Charlottetown, Prince Edward Island, Canada, 2000.

4. Syed, S. J. and A. M. Khan. Factor analysis for the study of determinants of public transit ridership, Journal of Public Transportation, Vol. 3, No. 3, 1-17, 2000.
5. Taylor, B. D. and C. N. Y. Fink. Explaining transit ridership: What has the evidence shown? Transportation Letters: the International Journal of Transportation Research, Vol. 5. No. 1, 15-26, 2013.
6. Githui, J. N., T. Okamura, and F. Nakamura. The structure of users' satisfaction on urban public transport service n developing country: the case of Nairobi," Journal of the Eastern Asia Society for Transport Studies, Vol. 8, pp. 1288-1300, 2010.
7. de Abreu e Silva, J., C. Morency, and G. G. Konstadinos. Using structural equations modeling to unravel the influence of land use patterns on travel behavior of workers in Montreal, Transportation Research Part A: Policy and Practice, Vol 46, No. 8, 1252-1264, 2012.
8. Kim, S. An Analysis of Elderly Mobility using Structural Equation Modeling, Transportation Research Board Annual Meeting, January 2003.
9. Mueller, R. Basic Principles of Structural Equation Modeling: An Introduction to LISTREL and EQS. Springer, New York, 1996.
10. Jöreskog, K.G.; Van Thillo, M. LISREL: A General Computer Program for Estimating a Linear Structural Equation System Involving Multiple Indicators of Unmeasured Variables (RB-72-56). Princeton, NJ: Educational Testing Service, 1972.
11. Lee, S-Y. Structural Equation Modeling, A Bayesian Approach, Wiley, New York, 2007.
12. National Transit Database Glossary, Federal Transit Administration, US Department of Transportation, September 2013, http://www.ntdprogram.gov/ntdprogram/Glossary.htm
13. Plummer, Martyn (2014). rjags: Bayesian graphical models using MCMC. R package version 3-12. http://CRAN.R-project.org/package=rjags
14. R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.