A QUALITY CONTROL FRAMEWORK FOR BUS SCHEDULE RELIABILITY

JIE LIN*, University of Illinois at Chicago
MING L. WANG, University of Illinois at Chicago
DAROLD T. BARNUM, University of Illinois at Chicago

*Corresponding Author

Great Cities Institute
College of Urban Planning and Public Affairs
University of Illinois at Chicago

Great Cities Institute Publication Number: GCP-08-06

A Great Cities Institute Working Paper

May 2008

The Great Cities Institute

The Great Cities Institute is an interdisciplinary, applied urban research unit within the College of Urban Planning and Public Affairs at the University of Illinois at Chicago (UIC). Its mission is to create, disseminate, and apply interdisciplinary knowledge on urban areas. Faculty from UIC and elsewhere work collaboratively on urban issues through interdisciplinary research, outreach and education projects.

About the Authors

JIE LIN*, University of Illinois at Chicago is a Professor in the Department of Civil and Materials Engineering & Institute for Environmental Science and Policy at the University of Illinois at Chicago.

MING L. WANG, University of Illinois at Chicago is a Professor in the Department of Civil and Materials Engineering at the University of Illinois at Chicago.

DAROLD T. BARNUM, University of Illinois at Chicago is Professor of Management and Professor of Information & Decision Sciences at the University of Illinois at Chicago.

Great Cities Institute Publication Number: GCP-08-06

The views expressed in this report represent those of the author(s) and not necessarily those of the Great Cities Institute or the University of Illinois at Chicago. GCI working papers may represent research in progress. Inclusion here does not preclude final preparation for publication elsewhere.

Great Cities Institute (MC 107)
College of Urban Planning and Public Affairs
University of Illinois at Chicago
412 S. Peoria Street, Suite 400
Chicago IL 60607-7067
Phone: 312-996-8700
Fax: 312-996-8933
http://www.uic.edu/cuppa/gci

A QUALITY CONTROL FRAMEWORK FOR BUS SCHEDULE RELIABILITY

Abstract

This paper develops and demonstrates a quality control framework for bus schedule reliability. Automatic Vehicle Location (AVL) devices provide necessary data; Data Envelopment Analysis (DEA) yields a valid summary measure from partial reliability indicators; and Panel Data Analysis provides statistical confidence boundaries for each route-direction's DEA scores. If a route-direction's most recent DEA score is below its lower boundary, it is identified as in need of immediate attention. The framework is applied to 29 weeks of AVL data from 24 Chicago Transit Authority bus routes (and therefore 48 route-directions), thereby demonstrating that it can provide quick and accurate quality control.

Key words: schedule adherence, data envelopment analysis (DEA), Panel Data Analysis (PDA), confidence interval, automatic vehicle location (AVL)

1 Introduction

The importance of reliable bus service to customers is well known, with “arriving when planned” being the most important desire of transit riders (Nakanishi, 1997; Transportation Research Board, 2002). Not surprisingly, consistency of service is one of the key sets of bus performance indicators that are monitored by most transit systems (Benn, 1995). Public transit agencies have developed multiple indicators to measure consistency of service, with indicators of on-time performance and headway adherence being almost universal, and a third common measure being running time adherence (Nakanishi, 1997; Benn, 1995; Vuchic, 2004; Transportation Research Board, 2003)

Unfortunately, the value of these service reliability indicators has been diminished by three problems. The first problem has been their infrequent collection. In order to make the best use of these indicators, it is necessary to frequently collect samples from each bus route, and to quickly make them available for analysis. In the past, for this activity to occur would have resulted in unacceptably high expenses because the data had to be collected and recorded manually (Nakanishi, 1997).

The second problem has been the absence of a single, over-all performance indicator that validly aggregates partial measures such as those identified above. One comprehensive service reliability indicator would make it much easier to quickly and validly identify those routes most in need of intervention. With multiple indicators, it may be difficult to determine which routes have the overall worst performance because routes doing well on some measures may be doing poorly on others. This problem is exacerbated when quick decisions should be made.

The third problem is determining whether a route’s declining service is caused by systematic new problems, or simply due to random chance. If management is to address problems of routes that are truly in difficulty, it should avoid wasting time on routes whose reported declines are simply random variations.

The purpose of this paper is to present a framework for mitigating these three problems, thereby enabling management to more quickly and accurately identify those routes most in need of assistance. The framework involves use of (1) Automatic Vehicle Location (AVL) data to obtain frequent and quickly-available samples, (2) Data Envelopment Analysis (DEA) to aggregate the various service reliability measures into one comprehensive indicator, and (3) Panel Data Analysis (PDA) to develop quality control charts for the performance of each individual route, which will alert management to routes performing worse than normal random variation explains.

The paper unfolds as follows. In the rest of this introductory section, background information on AVL, DEA and PDA is presented. Then, application of the framework is illustrated through a case study using archived AVL data provided by the Chicago Transit Authority (CTA). The CTA's bus route schedule adherence performance measures are defined in Section 2. The assessment framework is presented in Section 3. The case study results are reported in Section 4, including discussion on the DEA scores and their confidence intervals as quality controls for bus schedule adherence performance. Finally, the study contributions, limitations of the study and future research needs are summarized in Section 5.

1.1 Availability of Automatic Vehicle Location Data

With automatic vehicle location (AVL) devices becoming available on many buses in recent years, the quantity and quality of data have greatly improved and can be made quickly available to transit agencies. According to the U.S. Department of Transportation, two thirds of the 19 largest American transit agencies had their fleet fully equipped with AVL technology by 2004; the Chicago Transit Authority (CTA) is among those 100% AVL equipped agencies (U.S.

Department of Transportation, 2007). Therefore, AVL has become widespread and will likely be available at even more transit agencies in the future.

1.2 Data Envelopment Analysis

DEA is widely used in economic analysis for identifying technically efficient operations (Cooper et al., 2004; Färe et al., 1994; Färe and Grosskopf, 2004; Gattoufi et al., 2004). It is a linear programming method that combines partial efficiency measures into a single comprehensive indicator that provides objective evaluation and consistent comparisons of technical efficiency among decision making units (DMUs), i.e., base analysis units in DEA.

Use of DEA to compare the efficiencies of urban transit systems has become increasingly popular in recent years, particularly since 2000. De Borger et al. (2002) and Brons et al (2005) have given comprehensive reviews of transit DEA studies. Among the articles published since 2000, some analyze the efficiency of public transit in terms of services delivered (Graham, in press; Karlaftis, 2003&2004; Pina and Torres, 2001; Novaes, 2001); some measure the efficiency in terms of productivity (Odeck and Alkadi, 2001; Odeck, 2006); others compare technical and social efficiency of transit agencies (Boilé, 2001; Boame, 2004; Nolan et al., 2001&2002). One recent study uses Panel Data Analysis to make statistical inferences about estimated technical efficiencies of Canadian paratransit systems (Barnum et al., 2007a). Most recently, DEA has been applied to compare subunits within a single transit agency. Sheth et al. (2007) evaluated the overall performance of an agency's bus routes by using DEA and goal programming with artificial data. In one study, Barnum and his colleagues (2007b) combine DEA and Stochastic Frontier Analysis to compare the CTA's park-and-ride lot efficiency, and combine DEA and a reverse two-stage procedure (Barnum and Gleason, 2007) to analyze the technical efficiency of bus routes in another (Barnum et al., 2007c).

In this study, DEA scores are based solely on outputs (i.e., schedule adherence performance), which is different from all past transit DEA studies. In all of the other transit studies, DEA scores are based on output/input ratios known as technical efficiency indicators (Charnes et al., 1978; Cooper et al., 2004; Färe et al., 1994). The DEA scores in this study are effectiveness indicators, because they measure goal achievement (Gleason and Barnum, 1982). DEA has been so used in non-transit cases, such as in determining best location in location analysis (Thompson et al., 1986) and in evaluating human performance (Anderson and Sharp, 1997). In these two cases, as in this study, the DEA score is a measure of comparative output performance of each DMU, not a measure of each DMU's efficiency. Specifically, DEA is used in this study to compare the schedule adherence performance of individual routes.

1.3 Panel Data Analysis

Even with quickly available AVL data from which performance indicators can be aggregated into a single valid measure by DEA, a third problem remains to be solved. Because of random noise in the data (Grosskopf, 1996), a given route's sample DEA scores would be expected to vary; if a route's score goes down, or even if the trend in its scores is downward, this doesn't necessarily mean that the route's true performance has declined or is declining. In order to determine whether a decrease in a route's mean performance has occurred, it is necessary to develop statistical tests to determine if the scores have declined from their expected value to a statistically significant degree, or if the observed variations are just due to random chance. Recently, a new method has been developed using Panel Data Analysis on DEA scores to construct confidence intervals (Barnum et al., 2007a). This technique has not been applied to transportation subunits, but, because AVL makes panel data for individual bus routes available, it can be applied in this study.

2 CTA Bus Service Reliability Indicators

As the second largest transit agency in the nation, the CTA serves Chicago and forty surrounding suburbs with over 150 routes, more than 2,000 buses, and 2,273 route miles. CTA buses provide one million passenger trips a day and serve more than 12,000 posted bus stops.

The CTA has historically used running time adherence and headway regularity as key service reliability indicators (Hammerle et al, 2005). Not by coincidence, New York City Transit has applied the similar metrics (Nakanishi, 1997). Time-point level running time adherence and headway regularity are adopted in this study. They are calculated from the CTA's archived AVL data. In bus scheduling, time points are important physical points on a bus line that indicate when the bus is expected to arrive at those locations. In other words, buses are "timed" at time points rather than at stops, for example in CTA bus scheduling. Although stop level analysis could provide greater details of bus service, time point data satisfy the transit agency's practical needs and the purpose of this study for schedule adherence assessment, with much less data, storage requirements, and computational power.

2.1 Running time adherence

Running time adherence (measured in %) is defined as the average difference between the actual and the scheduled running times relative to the scheduled running time. When the actual running time is shorter than the schedule, the measure is called *Δ% Shorter Running Time* and otherwise *Δ% Longer Running Time*:

$$\Delta\%\ Shorter\ Running\ Time = \frac{\sum_{m}\left|\frac{Actual\ Run\ Time - Scheduled\ Run\ Time}{Scheduled\ Run\ Time}\right|}{m}\times 100\% \qquad (1)$$

$$\Delta\%\ Longer\ Running\ Time = \frac{\sum_{k}\frac{Actual\ Run\ Time - Scheduled\ Run\ Time}{Scheduled\ Run\ Time}}{k}\times 100\% \qquad (2)$$

Where *m* is the number of shorter running time events and *k* is the number of longer running time events between two consecutive time points in the same route-direction. The higher the running time metrics the worse the running time adherence.

2.2 Headway regularity

Similarly, headway regularity (measured in %) is defined as the average difference between the actual and the scheduled headways relative to the scheduled headway. If two consecutive buses are further from (or closer to) each other than the scheduled headway, the difference is called a longer (or shorter) headway difference. Bus bunching is an extreme example of short headway. The definition equations for headway regularity metrics are shown in Equations (3) and (4):

$$\Delta\%\ Longer\ Headway = \frac{\sum_{n} \frac{(Actual\ headway_i - Scheduled\ headway_i)}{Scheduled\ headway_i}}{n} \times 100\% \tag{3}$$

$$\Delta\%\ ShorterHeadway = \frac{\sum_{l} \left| \frac{(Actualheadway_j - Scheduledheadway_j)}{Scheduledheadway_j} \right|}{l} \times 100\% \tag{4}$$

Where *n* and *l* are the numbers of longer and shorter headway difference, respectively, at time points in the same route-direction. A high headway metric value indicates poor headway regularity adherence.

It is worth noting two bus operations/data phenomena that require particular attention in calculating headways with AVL data: bus overtaking and missing observations. Bus overtaking refers to the phenomenon in which the successor bus passes its predecessor along the route. From the view point of those waiting for a bus at stops, this makes no difference in measuring the headway, because the headway to them is always the time elapse between the last bus having arrived and the next bus to show up, whether the next bus is the scheduled bus or not. Therefore, in this paper, headway is calculated between the two sequential (in time) observations at a time point in the same route-direction.

Missing AVL data is not uncommon, occurring when the AVL device is broken or when tall buildings or other blockages (e.g., tunnels) exist to impair the ability of on-board GPS devices to determine location. In such cases the calculated headways are not the actual headways occurring in the field. On the other hand, the headway metrics defined in Equations (3) and (4) are relative measures between the actual and the scheduled headways. The absolute values of the observed headways are of little relevance if both the observed and the corresponding scheduled records are present in the dataset. For example, if bus trip No.2 is missing then both the actual and the scheduled headways become those between trips No.1 and No.3 and thus the headway difference is the sum of two pairs', No.1 and No.2, and No.2 and No.3. This calculation is consistent with the definitions in Equations (3) and (4).

3 Bus Schedule Adherence Assessment Framework

The quality control framework for bus schedule reliability consists of the following three steps. First, the DMU is determined to be bus route-direction in consideration of traffic directionality – the terms "route-direction" and "DMU" are interchangeable in the rest of the paper. The running time and headway metrics defined earlier are calculated for each bus route-direction. Second, DEA scores combining all four metrics are computed for each route-direction. These route-directions are then ranked by the DEA scores. Third, confidence intervals for the DEA scores and trends are computed for each route-direction, using PDA methodology (Kumbhakar and Lovell, 2000; Wooldridge, 2002; Hsiao, 2003; Frees, 2004; Baltagi, 2005; Baum, 2006). These confidence intervals serve as quality control limits for each route-direction's on-time performance. As demonstrated later in the paper, the findings of confidence interval analysis are important in interpretation of the DEA scores and have practical implications for control of bus on-time performance.

This framework is illustrated through a case study of twenty-four CTA key routes. These routes represent half of the CTA key routes (as opposed to support and special routes) and are spatially evenly distributed across the Chicago downtown and nearby suburbs. Ten of them are in the east-west direction, and the other fourteen are in the north-south direction. The 24 study routes are further divided into 48 route-directions (DMUs), labeled with letters. The eastbound and northbound route-directions are labeled with uppercase letters and their opposite westbound and southbound route-directions are labeled with the same but lowercase letter. For example, letters *B* and *b* represent the two opposite directions of the same bus route. The study time period covers weekday morning peak hours (6:30:00 AM to 8:59:59 AM) between January and June 2006, a total of twenty-nine weeks of AVL data. That is, a total of 48 x 29 = 1,392 route-directions for analysis.

3.1 Derivation of DEA-based Performance Measure

DEA uses linear programming techniques to weight and aggregate outputs divided by inputs in a way that results in a single comprehensive efficiency measure, with efficient units scoring exactly 100 (in percent) or beyond in the case of super-efficiency (Andersen and Petersen, 1993). The efficiency level of each non-efficient unit is expressed as a percentage of the efficiency of its efficient peers and is thus less than 100%.

In this study, the DEA model in Equation (5) is applied to measuring bus route-directions' schedule adherence performance, not their efficiencies. Note that although the four schedule adherence indicators are in truth outputs, they are treated as inputs in the DEA model. This is because DEA assumes increasing outputs are desirable and increasing inputs are undesirable. Because increases in those four indicators are undesirable, they are treated as inputs rather than outputs, a conventional method for dealing with undesirable outputs (Coelli et al., 2005). There are other ways to enter undesirable outputs into a DEA model (Scheel, 2001; Coelli et al., 2007), but discussion of them is beyond the scope of this paper.

$$\min_{\theta,\lambda} \theta$$

$$\begin{aligned} \text{Subject to:} \quad & \sum_{j=1}^{N} x_{jm}\lambda_j \le \theta x_{km} && m = 1,2,3,4 \\ & \sum_{j=1}^{N} y_j\lambda_j \ge y_k && y_j = y_k = 1, \forall j,k \\ & \lambda_k = 0 && \\ & \lambda_j \ge 0 && \forall j,\ j = 1,2,\ldots,1392,\ j \ne k \end{aligned} \qquad (5)$$

For each of the *j* route-directions (*j* = 1,…, 1392), there are four inputs, x_{jm}'s (*m* = 1,…,4), corresponding to the four schedule adherence indicators, i.e., Δ% *Shorter Running Time*, Δ% *Longer Running Time*, Δ% *Shorter Headway*, and Δ% *Longer Headway*; there is one output variable, y_j (*j* = 1,…, 1392), equal and set to unity for all route-directions.

Equation (5) is applied to each target route-direction *k* (k=1,…, 1392). Each time optimal weights λ_j's ($j \ne k$) are assigned to route-directions such that the target route-direction *k* receives the highest super-efficiency score *θ* (in %) it can possibly receive when compared to the other route-directions. The constraint $\lambda_k = 0$ prohibits the target route-direction *k* to be included as its own peer. Thus, if route-direction *k* is among the best performing ones, also known as the benchmark DMUs, its score is not limited to 100% as in the classical DEA efficiency scores but will reflect how much better performance it has than its benchmark peers. For example, if a route-direction has a score of 150%, then its score is 50% more than is needed to equal the other benchmark route-directions' performance. If a route-direction's score is less than 100%, it is still out-performed by benchmark route-directions even under its best possible performance.

Note that model (5) computes a common frontier for data from all weeks. This is called an intertemporal frontier. In fact, there are three different kinds of frontiers: intertemporal – observations are compared all at once across time periods, contemporaneous – observations are compared only with others in the same period, and sequential – observations are compared only with others in the same or earlier periods (Tulkens and Vanden Eeckaut, 1995).

3.2 Derivation of Panel DEA-based Confidence Intervals as Quality Control

Deriving DEA scores for bus schedule adherence performance is only the first step toward improving bus service reliability. In this study, it is of particular interest to establish confidence intervals as *quality control* limits so to quickly inform management when a given route's schedule adherence performance has worsened more than could be expected by random chance. By identifying only the routes with true problems, management can be more productive in use of their time by practicing "management by exception."

The confidence interval analysis is intended for quality control of bus schedule adherence performance. A route-direction requires immediate attention if it has at least one of the following three problems: (1) it is among those with the lowest performance scores; (2) its score for the most recent week is worse than the lower limit of its confidence interval; or (3) its performance scores show a statistically significant downward trend.

Although there will be situations in which the aforementioned problems cannot be corrected because of factors beyond transit agencies' control, the agencies will often be able to make changes that will improve the performance of problematic routes. Over time, the second two problems should become less common, although internal and external changes will always create new problems. There will always be routes "with the lowest performance scores." However, corrective actions over time should decrease the "depth" of the problem. It may be worth noting that improving the performance of the worst performers will not likely affect the future scores of other routes. All routes' performances are benchmarked to the best-performing routes, and it is highly unlikely that corrective actions will transform the worst performers into the best performers. Even if this were to happen, however, it is all to the good because it would increase the standards that all have to meet, which is a desirable situation.

PDA procedures (Kumbhakar and Lovell, 2000; Wooldridge, 2002; Hsiao, 2003; Frees, 2004; Baltagi, 2005) are used to derive confidence intervals of the DEA scores. The panel structure of the DEA scores, $P = \{\theta_{jt} \mid j = 1, 2, ..., 48;\ t = 1, 2, ..., 29\}$, makes it possible to use PDA, and therefore to estimate confidence intervals (Barnum et al., 2007a). A PDA approach represents a distinction from the pooled and cross-sectional methods that have been used in all published transit DEA studies. It also represents a desirable extension of the traditional DEA, which is a method of deterministic frontier analysis assuming that there is no statistical noise in the data and that scores measure efficiency without error.

A super-efficiency score θ_{jt} (for *t*=1,…, 29, and route-direction *j*=1,…, 48) includes a true "efficiency" value and a random error. Furthermore, for some of the routes, there may be a time trend in their performance[1]. Thus, the regression equation to be estimated can be written in the following form:

$$\theta_{jt} = \alpha_j + \beta_j z_t + u_{jt} \qquad j = 1,\ldots, 48;\ t = 1,\ldots,29 \qquad (6)$$

where θ_{jt} = the super-efficiency score for route-direction j and time t

α_j = schedule adherence performance score at *t*=1 associated with route-direction *j*,

[1] It would be worthwhile to identify influences such as environmental factors that cause the routes to have different on-time performance levels. However, the purpose of this paper is to present a methodology by which transit managers can use AVL data, DEA scores and PDA to construct control charts, so they can quickly react to routes that truly are in need of attention. It is not our attempt to empirically identify the reasons that the routes have different on-time performance levels, but valuable insight into the causes of performance differences among routes could be gained by such studies in the future.

β_j = coefficient of time trend for route-direction *j*, estimating the average weekly change in schedule adherence performance for that route-direction

z_t = week in sequence with values ranging from 0 to 28. The first week score is chosen as the base, i.e., z_1=0

u_{jt} = random error term for route-direction *j* in week *t.*

Super-efficiency scores are an observable proxy for latent variable values underlying conventional efficiency scores. Using super-efficiency scores avoids a limited-value response variable in the regression, as discussed by Coelli *et al* (2005). Thus, when using super-efficiency scores, it is not appropriate to use sample-selected, truncated or censored regression models, such as Tobit regression (Breen, 1996). Such methodologies would have been necessary if conventional technical efficiency scores, which yield a limited-value variable, had been used.

Equation (6) includes a variable, z_t, adjusting each route-direction's weekly performance. That is, the fitted value of the score of route-direction *j* is α_j in week 1, $\alpha_j + \beta_j$ in week 2, and so on up to $\alpha_j + 28\beta_j$ in week 29. This permits a different trend in the performance scores of each DMU, whether positive, negative or zero. If there is no change in on-time performance scores over time for route-direction *j*, or if the temporal trend is inconsistent, then β_j will not be statistically significant and α_j will be an unbiased estimate of the mean performance score of route-direction *j* for the entire 29-week period. If in truth the route-directions differ in their mean performance scores and/or the temporal trends, and if random errors are small compared to the true performance scores, Equation (6) should result in a statistically significant R-square.

For quality control purposes, scores above the upper confidence level indicate that performance has been better than expected and therefore do not require immediate corrective actions; scores falling below the lower confidence limit are those of concern. Therefore, only the lower limit of the confidence interval is estimated herein. The lower limit is $E(\theta_{jt}) - c_n s(\theta_{jt})$, where $E(\theta_{jt})$ is the expected value of θ_{jt}; c_n is the t-statistic at 0.10 level for a sample with *n* degrees of freedom; and $s(\theta_{jt})$ is the standard error of θ_{jt}, $s(\theta_{jt}) = \sqrt{MSE\left(1 + \mathbf{x}'_{jt}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{x}_{jt}\right)}$, where *MSE* is the sample's mean squared error; $\mathbf{x}_{jt}$ is the vector of regressor values used to predict the response variable θ_{jt}, and $\mathbf{X}$ is the matrix of independent variable values from the sample.

4 Case Study Results

4.1 Schedule Adherence Performance

Table 1 shows the descriptive statistics of the running time and headway metrics for the 48 route-directions over the 29 weeks of study period. On average, 41.34% of the actual running times are longer than (and thus behind) the scheduled, more than double of those shorter than (and thus ahead of) the scheduled (17.36%). For the headways, 43.46% are longer and 34.19% are shorter than the scheduled headways. The standard deviations of the longer running times and headways are also consistently larger than those of the shorter ones. These findings are expected, as buses running behind schedule are often the result of uncontrollable factors such as traffic congestion, whereas bus drivers have better control when running ahead of schedule.

Table 1. Descriptive statistics of running time and headway regularity metrics

Performance indicator	Minimum	Maximum	Mean	Std. Dev.
∆% Longer Running Time	13.58	158.46	41.34	20.40
∆% Shorter Running Time	6.36	31.04	17.36	4.48
∆% Longer Headway	3.88	145.10	43.46	21.63
∆% Shorter Headway	6.12	61.64	34.19	11.50

4.2 DEA-based Schedule Adherence Scores

The 29-week performance scores of the 48 route-directions show large variation across the route-directions and small variation within individual route-directions (Table 2). The analysis of variance (ANOVA) confirms the statistically, significantly different schedule adherence performances across route-directions (F-stat =186.567, $P(F_{47}$ >186.567)<0.000).

Table 2. Descriptive statistics of DEA scores

Route-direction	Mean	Variance	Route-direction	Mean	Variance
a	0.48	0.001	A	0.46	0.003
b	0.54	0.001	B	0.59	0.001
c	0.86	0.013	C	0.34	0.000
d	0.69	0.004	D	0.65	0.004
e	0.51	0.002	E	0.40	0.001
f	0.33	0.001	F	0.57	0.001
g	0.53	0.003	G	0.48	0.001
h	0.47	0.001	H	0.39	0.001
i	0.53	0.002	I	0.55	0.002
j	0.60	0.001	J	0.69	0.002
k	0.78	0.007	K	0.54	0.002
l	0.41	0.001	L	0.36	0.001
m	0.53	0.002	M	0.64	0.006
n	0.66	0.001	N	0.58	0.001
o	0.69	0.005	O	0.81	0.008
p	0.47	0.001	P	0.42	0.001
q	0.51	0.001	Q	0.71	0.004
r	0.64	0.001	R	0.53	0.002
s	0.45	0.001	S	0.65	0.007
t	0.54	0.001	T	0.52	0.001
u	0.59	0.001	U	0.33	0.000
v	0.73	0.004	V	0.69	0.005
w	0.46	0.000	W	0.34	0.000
x	0.56	0.011	X	0.78	0.014
Minimum = 0.33	Maximum = 0.86				

There are eight benchmark route-directions identified out of the 1,392 route-directions (Table 3). They are the ones with the DEA scores 100% or higher. These eight benchmarks come from four route-directions and three routes, *c, X*, and *O-o*, which had consistently good schedule adherence performance over the 29 weeks.

Table 3. Benchmark route-directions as a result of DEA

Route-Direction	Week	Score	Times of Being a Benchmark
c	1	104.13%	515
O	5	109.86%	186
c	8	101.09%	17
c	15	102.02%	24
X	27	120.64%	1354
O	30	194.39%	57
X	30	126.05%	672
o	30	117.08%	84

Most pairs' performance score distributions have quite different profiles. In particular, the inbound directions to downtown Chicago tend to have poorer schedule adherence in the morning rush hour. This directionality of traffic conditions justified the reason for separating route-directions for analysis.

It clearly is much easier to identify poor performance and trends when a valid summary measure such as a DEA score is used. However, the descriptive statistics of the DEA scores cannot identify either sudden decreases in performance below what could be expected by random chance, or downtrends in performance that are not just the result of random variation, both of which are of even more interest in practice. With Panel DEA scores computed from the archived AVL data, these issues can be addressed by combining the DEA scores with measures of statistical significance and confidence intervals for the DEA scores.

4.3 Estimates of PDA Parameters

The parameters of PDA (Equation 6) were estimated by using a fixed effects model and the Least-Squares Dummy Variables (LSDV) procedure (Baltagi, 2005). A fixed effects model was used rather than a random effects model or a mixed model because the intention was to estimate confidence intervals for the 48 route-directions (Baltagi, 2005). When the model includes both the 48 intercepts and the 48 weekly trend variables, 97 percent of the variance in DEA scores is explained by differences between the route-directions (R^2 = 0.9715, F = 464.65, P(F(95, 1296)>464.65) <0.00005)[2]. The intercepts (α_j's) and the slopes (β_j's) range between 0.020 and 0.296, and -0.0049 and 0.0068, respectively[3]. If the weekly trend variables are not included, R-square decreases from 0.97 to 0.85, with the difference between the full and reduced models being statistically significant ($\chi^2 = 260.85, P(\chi^2_{48} > 260.85) < 0.00005$).

The high R-square for the full model is not surprising, given the large differences between the route-directions' mean scores, and the small differences among the 29 scores of each route-direction, as discussed earlier (Table 2). Note that each of the 48 route-directions is an independent variable in Equation (6), and the trend in each of the 48 route-directions' schedule adherence performance is also an independent variable. So there are a total of 96 independent variables, two for each route-direction. The parameters of each of the 48 route-direction's pair of independent variables are estimated with 29 of the 1,392 (29 x 48) total observations. If the route-directions had not differed from each other systematically in their performance levels, then the R-square would have been very low; in this situation, they could be analyzed altogether with a single control chart. Likewise, if the route-directions' trends had not

[2] The error term in Equation (6) was corrected for heteroscedasticity using the same method applied in Barnum et al. (2007a). The error term was also found to have weak contemporaneous and serial correlation. The conservative decision adopted in our analysis was to assume the errors were independent. This was conservative because models accounting for error correlation are less robust and estimate narrower variances than i.i.d. variances. The error term distribution was slightly leptokurtic, although very close to normal in the tails; this should not affect our analysis results because it was the distribution tails that were of concern.

[3] The individual regression coefficients are not presented herein because they are of little interest other than to predict the confidence intervals described later. Another practical reason is the space limitation.

differed from each other systematically, then the R-square would not have improved when the full model replaced the reduced model.

4.4 Confidence Intervals for Schedule Adherence Performance

The confidence intervals of the performance scores for each route-direction and the statistical significance of trends were estimated based on the fixed-effect model of Equation (6). Table 4 identifies those routes most in need of action as of the final week of study: (i) those whose estimated performance scores for week 29 are in the thirty percent range, (ii) those whose actual performance scores for week 29 are below their expected value to a statistically significant degree (in the 0.10 tail), and (iii) those with a statistically significant (at the 0.05 level) downward trend performance over the 29-week period.

Table 4. Routes in need of management attention in week 29

DMU	Lowest Expected Scores*	Outside Lower CI*** (% below expected score)	Downward Trends** (Amount per Week)	Number of Problems
H	37.26%	-8.98%	-0.14%	3
f	36.22%	-7.65%		2
U	31.99%		-0.09%	2
A	39.72%		-0.46%	2
k		-10.63%		1
D		-9.24%		1
J		-8.10%		1
W	34.57%			1
C	34.66%			1
L	35.90%			1
c			-0.49%	1
i			-0.38%	1
g			-0.30%	1
r			-0.14%	1
a			-0.11%	1

Notes: *All DMUs in 29th week with expected score values in the 30% range.
**All DMUs with downward trends statistically significant at the 0.05 level.
***All DMUs whose actual scores in the 29th week were below the 0.90 confidence limits.

As can be seen, Route-direction *H* is the one most in need of attention. Its expected DEA score is among the lowest decile of scores; its scores have shown a statistically significant downward trend over the 29-week period; and, perhaps worst of all, its most recent score is below its expected score to a statistically significant degree, which may indicate an even steeper downtrend in the future. Three route-directions (*f, U* and *A*) are the next group of route-directions that demand attention, each having two problems, and the remaining poorly-performing routes each report one problem.

In some cases, poor performance may be the result of factors external to the transit agency, but in others corrections can and should be made. For example, among the problematic bus route-directions, *f* is a southbound service on a major arterial connecting several major intersections and CTA transfer points. The route also serves a medical district, where patient activities and traffic calming in the area are likely to slow down the bus service. Route-direction *W* is a bus route on the City's south side. There are seven high schools, two Metra[4] train stations and one CTA train station on this line, which may have contributed to its low scores. Route-direction *H* has a similar story to *W*, but clearly has more performance issues.

The route-direction pair *C* and *c* have contrasting performance, where *c* generally performs well (see Table 2). The bus line runs north-south (*C*-*c*) between the city's far south side and downtown. In the morning, the northbound (inbound) buses encounter much higher passenger activities and traffic going into the city. On the other hand, our confidence interval analysis indicates that the southbound (*c*) is not problem-free either – although it has high performance as of week 29, its long-term performance is trending down.

[4] The 495-mile Metra system serves 230 stations between suburban Chicago and the City of Chicago

As a comparison, if the four performance indicators were used directly to assess the routes' schedule adherence, Table 5 lists the top five worst route-directions for each of the four performance indicators. The ranking is based on the mean DEA scores over the 29 week period. Three points are observed. First, it is not obvious as to which route-direction has the worst performance overall. A route-direction may perform reasonably well by one measure but poorly by the others. Second, eight out of the fifteen problematic route-directions identified by confidence interval analysis (in Table 4) are not detected using only the historical averages of the individual performance indictors (see Table 5). In particular, Route-direction *H*, identified as the most in need of attention in confidence interval analysis, is not in the top five in any of the four lists in Table 5. Lastly, downward trends are not directly obvious when the individual indicators are used. For example, route-direction *c* performs generally well and thus is not seen on any of the four lists, but has been identified with a significant downward trend by PDA.

Table 5. Top five worst performing route-directions by individual performance indicators

Route-	**Δ% Longer Running Time**		**Route-**	**Δ% Shorter Running Time**	
direction	**Mean**	**Variance**	**direction**	**Mean**	**Variance**
P	113.87	177.11	F	28.51	0.36
h	104.28	251.09	U	26.94	0.93
C	80.41	205.11	f	26.53	3.03
I	73.33	53.74	A	23.57	0.95
L	70.67	110.48	E	22.92	1.59
Route-	**Δ% Longer Headway**		**Route-**	**Δ% Shorter Headway**	
direction	**Mean**	**Variance**	**direction**	**Mean**	**Variance**
W	105.90	257.35	w	55.21	7.81
w	99.00	324.19	W	54.83	15.37
a	81.17	146.69	j	53.30	6.94
j	78.71	94.08	a	51.56	11.59
S	74.77	181.00	L	47.42	25.91

The above findings re-emphasize two important points. First, validly combining a variety of performance measures into a single indicator makes it much easier to identify problem routes. If the four original indicators had been used instead of the one summary measure, identifying those routes most in need of action would have been very difficult if not impossible. Second, use of statistical significance allows management to concentrate on those routes that truly are in trouble and to avoid taking action on those routes whose scores are within the range of normal random variation.

5 Summary and Conclusions

This study presented a Panel DEA framework for evaluating bus schedule adherence performance. The proposed framework was demonstrated with a case study using twenty-four CTA bus routes for a twenty-nine-week period between January and June 2006. The bus schedule adherence performance indicators, namely running time adherence and headway regularity, were derived from CTA's archived bus AVL data. Compared to assessing bus schedule adherence based on partial performance indicators historically used by transit agencies, the Panel DEA-based framework demonstrates clear superiority in terms of providing a comprehensive performance measure that identifies problems quickly and accurately.

The contributions of the paper to the transit literature are four-fold. First, this paper has demonstrated a new application of AVL data for transit operations. Because AVL data is continuously collected and quickly available, management can use the information to promptly address service reliability problems. Moreover, trend analysis and panel data analysis become practical when AVL data are available.

Second, this paper presents a mathematically and economically plausible method to construct a comprehensive measure of service reliability from multiple partial reliability indicators, by using DEA. This DEA indicator is put into even better use than those in all past transit DEA studies because it not only identifies the benchmark DMUs but also prioritizes those in need of attention. Prior transit DEA studies have usually addressed the former but not the latter. In addition, the DEA indicator developed herein is a pure effectiveness measure in that it utilizes only outputs, instead of a ratio of outputs to inputs that have been employed in all past transit DEA studies. It is likely that there are other transit goals that could be best measured by considering only outputs rather than output-input ratios.

Third, by coupling PDA with DEA, it has been demonstrated in this paper that the traditional deterministic DEA can be extended to stochastic DEA and thus statistical inferences can be made. This is the second application of Panel DEA to transit and the first to apply the methodology to performance evaluation of a transit agency's subunits. Unlike all previous transit DEA papers, this paper is built upon the concept of quality control, realized by comparing each DMU's current performance level to a statistical confidence interval based on its past performance.

Lastly, this paper is among the first that have extended the use of DEA from the traditional comparisons among transit organizations to performance assessment of organizational subunits performing parallel activities.

Nonetheless, there are limitations in this study that require further research. It is recognized that the four on-time performance indicators used in this study do not reflect every aspect of bus service reliability. For example, no measure of passenger related activity was considered. Nor have measures of traffic conditions or environmental factors been taken into consideration. The running time and headway based indicators were adopted mainly because they have been used by transit agencies (e.g., CTA) and they were readily derived from the archived AVL data. Some of the other measures may be derived from the bus automatic passenger count (APC) data; others may require different data sources such as traffic sensor data. The real-time AVL data is another source of data, which contain travel speed information. With more advanced data collection technologies available in public transportation systems, the framework presented in the paper should readily apply.

Future research is also needed to identify causes for poor bus schedule reliability. This is outside the scope of this paper but no-doubt a close-to-the-heart issue to transit managers. Panel DEA tells which routes demand immediate attention; however, it does not identify the causes of the problems. Continuous research effort in the area is desired and will surely elevate the value of the research even further.

6 Acknowledgements

This project was partially supported by the Chicago Transit Authority under contract to the Urban Transportation Center at the University of Illinois at Chicago, which made this research possible. We especially thank Michael Haynes of CTA for his help in providing AVL data for this research. Finally, we thank the three reviewers and the editor whose wise counsel greatly improved the paper. The views expressed in the paper are those of the authors alone.

7 References

Andersen, P. and Petersen, N.C. (1993) A procedure for ranking efficient units in data envelopment analysis. *Management Science*, 39 (10), 1261-1265.

Anderson, T.R., and Sharp, G.P. (1997) A New Measure of Baseball Batters Using DEA. *Annals of Operations Research*, Vol. 73, No. 0, 141-155

Baltagi, B. H. (2005) *Econometric Analysis of Panel Data*. John Wiley & Sons, Ltd, West Sussex, England.

Barnum, D.T., Gleason, J.M., and Hemily, B. (2007a) Estimating Confidence Intervals for DEA Scores with Panel Data Analysis. Forthcoming in *Applied Economics.*

Barnum, D.T., McNeil, S., and Hart, J. (2007b) Comparing the efficiency of public transportation subunits using data envelopment analysis. *Journal of Public Transportation*, 10 (2), 1-16.
Barnum, D. T., Tandon, S. and McNeil, S. (2007c) Comparing the performance of bus routes after adjusting for the environment, using data envelopment analysis. Forthcoming in *Journal of Transportation Engineering*.

Barnum, D.T. and Gleason, J.M. (2007) Bias and precision in the DEA two-stage method. Forthcoming in *Applied Economics*.

Baum, C.F. (2006) *An Introduction to Modern Econometrics Using Stata*. Stata Press, College Station, Texas

Benn, H.P. (1995) *Bus Route Evaluation Standards, Transit Cooperative Research Program, Synthesis of Transit Practice 10*. National Academies, Washington, DC

Boame, A.K. (2004) The Technical Efficiency of Canadian Urban Transit Systems. *Transportation Research Part E-Logistics and Transportation Review*, Vol. 40, No. 5, pp. 401-416

Boilé, M.P. (2001) Estimating Technical and Scale Inefficiencies of Public Transit Systems. *Journal of Transportation Engineering*, Vol. 127, No. 3, pp. 187-194

Breen, R. (1996) *Regression Models: Censored, Sample-Selected, or Truncated Data*. Sage Publications, Thousand Oaks, Ca.

Brons, M., Nijkamp, P., Pels, E., and Rietveld, P. (2005) Efficiency of Urban Public Transit: A Meta Analysis. *Transportation*, Vol. 32, No. 1, pp. 1-21

Charnes, A., Cooper, W.W., and Rhodes, E. (1978) Measuring the efficiency of decision making units. *European Journal of Operational Research*, Vol. 2, No.6, pp. 429-444

Coelli, T. J., Lauwers, L. and Van Huylenbroeck, G. (2007) Environmental efficiency measurement and the materials balance condition. Forthcoming in *Journal of Productivity Analysis*.

Coelli, T.J., Rao, D.S.P., O'Donnell, C.J., and Battese, G.E. (2005) *An Introduction to Efficiency and Productivity Analysis*. Springer, New York, NY.

Cooper, W.W., Seiford, L.M., and Zhu, J. (2004) *Handbook on Data Envelopment Analysis*. Kluwer Academic Publishers, Boston, MA

De Borger, B., Kerstens, K., and Costa, A. (2002) Public Transit Performance: What Does One Learn From Frontier Studies? *Transport Reviews*, Vol. 22, No. 1, pp. 1-38

Färe, R., Grosskopf, S., and Lovell, C.A.K. (1994) *Production Frontiers*. Cambridge University Press, Cambridge, England

Färe, R. and Grosskopf, S. (2004) *New Directions: Efficiency and Productivity*. Kluwer Academic Publishers, Boston

Frees, E.W. (2004) *Longitudinal and Panel Data: Analysis and Applications in the Social Sciences*. Cambridge University Press, Cambridge, U.K.

Gattoufi, S., Oral, M. and Reisman, A. (2004) Data envelopment analysis literature: a bibliography update (1951-2001). *Socio-Economic Planning Sciences*, 38 159-229

Gleason, J.M. and Barnum, D.T. (1982). Toward valid measures of public sector productivity: Performance indicators in urban transit. *Management Science*, 28 (4), 379-386.

Graham, D.J. (in press) Productivity and Efficiency in Urban Railways: Parametric and Non-Parametric Estimates. *Transportation Research Part E: Logistics and Transportation Review*, online version available.

Grosskopf, S. (1996). Statistical inference and nonparametric efficiency: A selective survey, Journal of Productivity Analysis, 7(2-3), 161-176.

Hammerle, M., Haynes, M., McNeil, S. (2005) Use of Automatic Vehicle Location and Automatic Passenger Counter Data to evaluate bus operations for the Chicago Transit Authority. *Transportation Research Record* 1903, pp 27-34, Transportation Research Board, National Research Council, Washington D.C.

Hsiao, C. (2003) *Analysis of Panel Data*. Cambridge University Press, Cambridge.

Karlaftis, M.G. (2003) Investigating Transit Production and Performance: A Programming Approach. *Transportation Research Part A-Policy and Practice*, Vol. 37, No. 3, pp. 225-240

Karlaftis, M.G. (2004). A DEA approach for evaluating the efficiency and effectiveness to urban transit systems. *European Journal of Operational Research*, Vol. 152, No. 2, pp. 354-364.

Kumbhakar, S. and Lovell, C. A. K. (2000) *Stochastic Frontier Analysis*. Cambridge University Press, Cambridge England.

Nakanishi, Y.J. (1997) Bus Performance Indicators: on-Time Performance and Service Regularity. *Transportation Research Record*, Vol. 1571, pp. 3-13

Nolan, J.F., Ritchie, P.C., and Rowcroft, J.R. (2001) Measuring Efficiency in the Public Sector Using Nonparametric Frontier Estimators: A Study of Transit Agencies in the USA. *Applied Economics*, Vol. 33, No. 7, pp. 913-922

Nolan, J.F., Ritchie, P.C., and Rowcroft, J.E. (2002) Identifying and Measuring Public Policy Goals: ISTEA and the US Bus Transit Industry. *Journal of Economic Behavior & Organization*, Vol. 48, No. 3, pp. 291-304

Novaes, A.G.N. (2001) Rapid Transit Efficiency Analysis With the Assurance-Region DEA Method. *Pesquisa Operacional*, Vol. 21, No. 2, pp. 179-197

Odeck, J., and Alkadi, A. (2001) Evaluating Efficiency in the Norwegian Bus Industry Using Data Envelopment Analysis. *Transportation*, Vol. 28, No. 3, pp. 211-232

Odeck, J. (2006) Congestion, Ownership, Region of Operation, and Scale: Their Impact on Bus Operator Performance in Norway. *Socio-Economic Planning Sciences*, Vol. 40, No. 1, pp. 52-69

Pina, V., and Torres, L. (2001) Analysis of the Efficiency of Local Government Services Delivery: An Application to Urban Public Transport. *Transportation Research Part A-Policy and Practice*, Vol. 35, No. 10, pp. 929-944

Scheel, H. (2001) Undesirable outputs in efficiency valuations. *European Journal of Operational Research*, 132 (2), 400-410.

Sheth, C., Triantis, K., and Teodorovic, D. (2007) Performance evaluation of bus route: A provider and passenger perspective. *Transportation Research Part E: Logistics and Transportation Review*, Volume 43, Issue 4, pp. 453-478.

Thompson, R.G., Singleton, F.D., Thrall, R.M., and Smith, B.A. (1986) Comparative Site Evaluations for Locating A High-Energy Physics Lab in Texas. *Interfaces*, Vol. 16, No. 6, pp. 35-49

Transportation Research Board (2002) *Customer-Focused Transit*. A Transit Cooperative Research Program Synthesis 45, National Research Council, Washington, D.C.

Transportation Research Board (2003) *A Guidebook for Developing a Transit Performance-Measurement System.* A Transit Cooperative Research Program (TCRP) Report No. 88, National Research Council, Washington, D.C.

Tulkens, H. and Vanden Eeckaut, P. (1995) Nonparametric efficiency, progress and regress measures for panel-data - methodological aspects. *European Journal of Operational Research*, 80 (3), 474-499

U.S. Department of Transportation (2007) *ITS Deployment Statistics – transit buses with automatic vehicle location (AVL) and computer aided dispatch (CAD)*, available at http://www.itsdeployment.its.dot.gov/, last modified May 2007

Vuchic, V.R. (2004) *Urban Transit: Operations, Planning, and Economics*. Wiley, New York, NY
Wooldridge, J.M. (2002) *Econometric Analysis of Cross Section and Panel Data*. MIT Press, Cambridge, MA.