Estimating DEA Confidence Intervals with Statistical Panel Data Analysis

Darold T Barnum[a] John M Gleason[b] Matthew G. Karlaftis[c] Glen T Schumock[d] Karen L Shields[e] Sonali Tandon[f] Surrey M Walton[g]

[a](Corresponding author) Departments of Information & Decision Sciences, Managerial Studies, and Pharmacy Administration, University of Illinois at Chicago (MC 243), 601 South Morgan Street, Chicago, IL 60607-7123 USA, Phone: +1-312-996-3073, Fax: +1-312-996-3559, dbarnum@uic.edu

[b]Department of Information Systems and Technology, College of Business Administration, Creighton University, Omaha, NE 68178 USA, jgleason@creighton.edu

[c]Department of Transportation Planning Engineering, School of Civil Engineering, National Technical University of Athens, 5, Iroon Polytechniou Street, Zografou Campus, Athens 15773, Greece, mgk@central.ntua.gr

[d]Center for Pharmacoeconomic Research, and Department of Pharmacy Practice, College of Pharmacy, University of Illinois at Chicago (MC 886), Chicago, IL 60607 USA, schumock@uic.edu

[e]Sisters of St. Francis Health Services, Inc., 1515 W Dragoon Trl, Mishawaka, IN 46544-4710 USA, karen.shields@ssfhs.org

[f]Chicago Transit Authority, 567 W. Lake Street, Chicago IL 60661 USA, standon@transitchicago.com

[g]Center for Pharmacoeconomic Research, and Department of Pharmacy Administration, College of Pharmacy, University of Illinois at Chicago (MC 871), Chicago, IL 60607 USA, walton@uic.edu

***Abstract*:** This article describes a statistical method for estimating Data Envelopment Analysis (DEA) score confidence intervals for individual organizations or other entities. The method applies statistical Panel Data Analysis (PDA), which provides proven and powerful methodologies for diagnostic testing and for estimation of confidence intervals. DEA scores are tested for violations of the standard statistical assumptions including contemporaneous correlation, serial correlation, heteroskedasticity, and the absence of a Normal distribution. Generalized Least Squares statistical models are used to adjust for violations that are present, and to estimate valid confidence intervals within which the true efficiency of each individual decision making unit occurs. The method is illustrated with two sets of panel data, one from large U.S. urban transit systems and the other from a group of U.S. hospital pharmacies.

Key words: Data Envelopment Analysis, DEA, Econometrics, Efficiency, Probability, Statistics

Estimating DEA Confidence Intervals with Statistical Panel Data Analysis

1. Introduction

Developing statistical methodologies to deal with noise has become a significant focus of Data Envelopment Analysis (DEA) research. Chambers and Färe [12, p. 329] observe that "more and more effort has been devoted to determining the statistical properties of the DEA approach." It has been shown that DEA scores possess stochastic characteristics that permit many types of statistical estimations [3, 25]. Stochastic variations in DEA scores have been addressed by methodologies such as chance-constrained programming [15, 44], window analysis [17], sensitivity-robustness-stability analysis [16], bootstrapping [1, 28, 35-39], and, most recently, a promising new Bayesian procedure [22]. However, none of these methodologies can estimate, with a specified probability, the confidence interval for the true efficiency of an individual organization or other entity, herein called a Decision Making Unit (DMU).

Bootstrapping as developed to date does not consider the stochastic variations of individual DMUs (with the exception of Atkinson and Wilson's 1995 article [1], which has not been utilized since). Indeed, Simar and Wilson's seminal articles on bootstrapping specifically state that their methodology treats the input-output set of each DMU as constant [35, 36, 38]. So, bootstrapped confidence intervals estimate the range within which a fixed input-output set's true efficiency occurs with a specified probability, using the set of inputs and outputs from a particular DMU. But, the bootstrapping methodology does not incorporate stochastic variations in each individual DMU's output/input performance. Therefore, bootstrapping overestimates (by an unknown amount) the probability that a DMU's true efficiency will occur within the reported confidence limits. Further, because the stochastic variation within each DMU is frequently heteroskedastic across DMUs, the aforementioned unknown overestimation will vary by unknown amounts among DMUs.

Gajewski, Lee, Bott, Piamjariyakul and Taunton [22] have recently developed an innovative Bayesian methodology for estimating the best-practice frontier that represents an elegant alternative to bootstrapping. In general it requires fewer assumptions than does bootstrapping, and, more importantly, it introduces Bayesian statistical methodology to the DEA field. Given that bootstrapping and the new

Bayesian approach are based on different assumptions and statistical methodologies, researchers estimating best practice frontiers may wish to triangulate their estimates by utilizing both methods. However, the Bayesian methodology to date is also based on variations in the frontier while ignoring variations within individual DMUs. So it, like bootstrapping, it can estimate the probably distribution for the efficiency of a fixed set of inputs and outputs, but it cannot estimate the probability distribution for the efficiency of individual DMUs.

To date, both bootstrapping and Bayesian estimation have been based on only one observation of each DMU. It clearly is impossible to estimate an individual DMU's variation without multiple observations. So, for those circumstances where only cross-sectional data are available, bootstrapping and Bayesian estimation may be the best alternatives.

But, those relying on these methodologies cannot determine whether an assessed DMU is efficient (or inefficient) with a specified degree of statistical significance, nor construct a confidence interval within which the DMU's true efficiency will occur with a specified probability. Further, it is not possible to determine if an individual DMU's efficiency uptrends or downtrends are statistically significant or just random variations, nor whether ongoing processes evaluated by DEA scores are in or out of control.

2. Use of PDA with DEA, and the contribution of this paper

When multiple observations of each DMU's scores are available, as occurs with panel or clustered data, an expected mean score for each DMU can be computed. Then, we can use the distribution of the actual scores around the means to characterize the nature of the stochastic disturbances. All stochastic disturbances, including DEA score errors, can encapsulate complicated, unidentified interactions with other variables. Such disturbances can be treated as random if they satisfy appropriate empirical tests [21, 23, 24, 47]. Therefore, we can empirically test whether DEA score residuals are independent and identically distributed (i.i.d.) and Normally distributed. Statistical panel data analysis (PDA) methodologies can be used to identify violations of these requirements, and, where violations exist, to employ appropriate statistical models to correct for them in confidence interval estimations.

The first article using PDA with DEA to estimate efficiency confidence intervals for individual DMUs was published in the *Journal of Transportation Engineering* [5] in 2008. It estimated confidence intervals for the mean DEA scores of Canadian paratransit systems, with the data adjusted for environmental variations. PDA/DEA methodologies have since been utilized by articles in journals from several fields. These consist of articles identifying confidence intervals of Canadian paratransit system efficiencies that had not been adjusted for environment differences [6], replacing decision making under uncertainty with decision making under risk in operations research [7], identifying statistically-significant trends in the efficiencies of hospital pharmacies [9] and measuring bus schedule reliability (rather than efficiency) for individual routes with a DEA-inspired linear programming model [31].

This paper contributes to the body knowledge in several important respects. All of the preceding PDA/DEA papers have been directed to specific fields or industries, and all are published in journals whose target audiences are neither applied statisticians nor general users of applied statistical techniques. Because the PDA/DEA methodology would be useful across a wide range of DEA issues amenable to statistical applications, we feel it should be introduced to a broad audience of applied statisticians. In this paper we particularly emphasize the statistical methodology and assumptions, and their consequences for valid estimation. Because the PDA/DEA methodology uses parametric statistical models to estimate, validly for the first time, DEA score confidence intervals for individual units, it opens up a new application of such models to statisticians.

Further, the *Journal of Applied Statistics* is an especially appropriate outlet for presenting this methodology to applied statisticians because of *JAS*'s publication of other statistical methodologies for dealing with DEA scores. The first is one of the seminal articles on bootstrapping methodology [37] published in 2000, and another is the promising new Bayesian methodology published in 2009 [22]. Unfortunately, the 2000 DEA bootstrapping article [37] explicitly estimates DEA efficiency confidence intervals of individual DMUs, in its case schools. It therefore disseminates bootstrapping's fatal flaw of estimating DMU confidence intervals based solely on efficiency variations of the frontier while ignoring efficiency variations of individual DMUs. Based on this and several other key articles, bootstrapping

DEA efficiency scores to (incorrectly) develop confidence intervals for individual DMUs has become widespread since 2000. So, it seems appropriate that a valid alternative be published in this journal.

Finally, there is one further concern with bootstrapping as described in [37]. The authors make five assumptions that serve to characterize their Data Generating Process (DGP). Only the first would be subject to empirical testing, and in fact all five assumptions are accepted with no empirical evidence that they are valid for either the dataset analyzed in [37] or indeed for any other specific dataset. Although we have not addressed it in our early articles, the characteristics of the DGP assumed for PDA/DEA modeling can be empirically tested, as we specifically discuss in Section 5, with other heretofore overlooked statistical consequences identified in Section 9. These expositions should remind applied statisticians and others of the importance of both identifying and empirically validating DGP assumptions.

3. Organization of this paper

In this paper, we exhibit the use of PDA with two sets of panel data. One panel is DEA scores from 50 large urban transit systems, with 5 scores for each system. The other panel is DEA scores from 12 hospital pharmacies, with 13 scores for each pharmacy. Where heteroskedasticity, serial correlation and contemporaneous correlation are present, we identify Generalized Least Squares (GLS) statistical models that account for these conditions in their estimations. Using our two data sets, one that complies with the i.i.d. assumptions and the other that does not, we demonstrate how PDA methods can be used to estimate valid confidence intervals for the true efficiencies of individual DMUs.

First, we discuss our DEA model, the data generating process, and our statistical diagnostic tests. Next we apply the procedure to our pharmacy data: we describe the inputs and outputs, the PDA statistical model, the results of tests for i.i.d. and Normality, confidence intervals for each pharmacy's expected efficiency over time, and control charts for determining when a pharmacy's efficiency is "out of control." Then we apply the procedure to our transit data: we identify the inputs and outputs, the PDA statistical model, the results of tests for i.i.d. and Normality, and the confidence interval estimates for each individual transit system's efficiency. Finally, we discuss related issues and conclude.

4. The DEA model

The DEA scores are based on Linear Program 1. We utilize the Charnes-Cooper-Rhodes (CCR) input-oriented DEA model [13] adapted so its scores are not censored at 1. That is, instead of censoring input-oriented efficiency scores θ at one $(0 \le \theta \le 1)$, the model allows θ to vary over [0, ∞), where $\theta < 1$ is inefficient and $\theta \ge 1$ is efficient, by adding equation 1.3 to the conventional CCR model. For the *j* DMUs $(j = 1,...,J)$, there are data on *m* inputs $(x_{11},...,x_{jM})$, and on *n* outputs $(y_{11},...,y_{jN})$. The DEA score θ identifies the technical efficiency of the assessed DMU *k*. All DEAs are conducted with Scheel's EMS software [33]. We use this model with both the transit and the hospital pharmacy data.

$$\min_{\lambda} \theta \qquad (1.0)$$

subject to

$$\sum_{j=1}^{J} y_{jn}\lambda_j \ge y_{kn} \qquad n = 1,...,N \qquad (1.1)$$

$$\sum_{j=1}^{J} x_{jm}\lambda_j \le \theta x_{km} \qquad m = 1,...,M \qquad (1.2)$$

$$\lambda_k = 0 \qquad (1.3)$$

$$\lambda_j \ge 0 \qquad j = 1,...,J \qquad (1.4)$$

.

We use the CCR model because it has been demonstrated in several other studies of hospital pharmacies [26, 34], and confirmed for our data herein, that hospital pharmacies have constant returns to scale. Although transit systems sometimes produce decreasing returns to scale, our sample of large urban transit systems showed constant returns to scale, so the CCR model is also appropriate for that dataset.

The model is based on a contemporaneous frontier [45], that is, efficiencies are computed separately for each cross section of the data. We use a contemporaneous frontier in this paper, where efficiencies in each time period are estimated based on the most efficient DMUs in that time period, because it results in estimates more sensitive to the presence of contemporaneous correlation.

The conventional CCR model reports censored scores. The output/input ratio of the assessed DMU is compared to the output/input ratio of its efficient peers only if the assessed DMU is inefficient. If however the assessed DMU is efficient, then its output/input ratio is compared to its own output/input ratio, so its score cannot exceed 1.

Uncensored DEA scores result when the output/input ratio of the assessed DMU is compared to the output/input ratio of efficient peer DMUs, regardless of the assessed DMU's level of efficiency. (Note that equation 1.3 prevents the DMU being assessed from entering into the comparison base.)

For example, suppose the maximum value of an assessed DMU's aggregated and weighted outputs to aggregated and weighted inputs is 4/10, while its composite efficient peer (using the identical set of weights) has an output/input ratio of 5/10. In this case, because the assessed DMU is inefficient, the CCR model would report its efficiency as 0.8 whether or not the scores are censored.

Now, suppose that, in periods 2, 3 and 4, the assessed DMU's ratio is 5/10, 7/10 and 6/10, while the composite to which it is compared remains at 5/10. If the scores are censored, the assessed DMU's reported efficiencies for the three periods will be: 1.0, 1.0 and 1.0. If the scores are uncensored, then the assessed DMU's reported efficiencies will be: 1.0, 1.4 and 1.2.

From the viewpoint of deterministic estimation of efficiency in which the data are assumed to be non-stochastic, the censored score is sufficient and can serve as a direct proxy for efficiency, thereby following the DEA convention that relative efficiency can never be greater than 100 percent. One purpose of the CCR program is to identify DMUs that are on the production frontier, and a score of 1 symbolizes a point on the production frontier.

But, for statistical estimation, when ubiquitous stochastic variation is taken into account, the uncensored score is superior. Among other reasons, information useful for estimating statistical significance and statistical confidence is not discarded.

Utilizing the full range of scores $(0 \le \theta < \infty)$ does not affect which DMUs are reported to be efficient, nor does it affect the scores of any inefficient DMU. For any DMU with a score of 1 or greater, we know that no other DMU is more efficient at its location on the frontier. So, the estimated efficient frontier and the DMUs defining it will be identical whether or not scores are truncated at 1.

Likewise, the reported efficiencies of inefficient DMUs are not affected, because, whether or not a benchmark DMU's score is recorded as 1 or some higher value, the outputs and inputs underlying that score will be what determines the score of an inefficient DMU that it is benchmarking. For example,

suppose the DMU assessed above is a benchmark DMU for periods 2, 3 and 4. And, the DMU that it is benchmarking has a constant output/input ratio of 4/10 for all three periods. This constant ratio will be compared to the benchmark DMU's ratios of 5/10, 7/10 and 6/10, resulting in efficiency estimates of 0.8, 0.57 and 0.67 for the assessed DMU. This is true regardless of whether the benchmark's reported efficiencies are truncated at 1 for all three periods or reported as 1.0, 1.4 and 1.2. In short, using uncensored DEA scores provides valuable information for statistical estimation but has no effect on which DMUs are reported as efficient or on the efficiency scores of inefficient DMUs.

5. The data generating process

The appropriate method for creating confidence intervals depends on the characteristics of the data generating process (DGP). We do not assume that the DGP yields errors in DMU scores that are i.i.d. and Normally distributed, or that it does not. Rather, we hypothesize that deviations from (DMU level) mean scores are i.i.d. and Normally distributed. We empirically test the data, via examination of residuals around their DMU-specific means, to verify or reject our hypotheses. Then, we use methods for generating probabilities that have been justified by empirical evidence about the DGP rather than basing our choices on implicit or explicit assumptions.

As is true for traditional DEA [13, 17, 22], the DMUs in our datasets are the population, and the data cover the time period of concern. That is, we adopt the convention that a DMU's relative efficiency is determined solely by comparisons with the other real DMUs in the analysis. In many cases, those concerned with the performance of particular DMUs want to know how those DMUs compare with their actual competitors/peers. This certainly is true our pharmacy case: the reason for the control system is to compare the twelve pharmacies with each other and with themselves over time. In transit, policy and decision makers usually are most interested in how specific operations compare with other specific operations [29, 32, 40]. (If indeed the sample at hand does not represent the entire population of interest, then, building on current work [22], panel-data Bayesian methodologies that estimate the best practice frontier for the population would be a welcome addition to the literature.)

6. Statistical diagnostic testing

The standard assumptions are that the random errors in the DMU's efficiency scores are independent and identically distributed (i.i.d.), and Normally distributed. Before any attempt to construct significance levels or confidence intervals can be validly performed, it is necessary to confirm that the preceding assumptions hold. If the conditions do not hold, appropriate Generalized Least Squares (GLS) models would be necessary to account for violations. That is, one must empirically test the data to determine whether the DGP has produced DEA scores that are (or are not) i.i.d. and Normally distributed, and then use models justified by the evidence. All statistical analyses in this paper use Stata 10 [43].

To test for *Normal distribution* of the errors, we use the Shapiro-Wilk W test, the Shapiro-Francia W' test, and a joint skewness and kurtosis test [41]. To determine if the errors are *identically distributed*, we test for heteroskedasticity across the DMUs with the Breusch-Pagan/Cook-Weisberg method[24].

The errors are not *independent* if either serial or contemporaneous correlation is present. Serial correlation, often called autocorrelation, occurs among a DMU's error terms when its error in one time period is correlated with its errors in other time periods [2, 24, 30, 47]. We test for serial correlation using the Wooldridge test for autocorrelation in panel data [47]. Contemporaneous correlation, also called cross-sectional correlation, cross-sectional dependence, and spatial correlation, occurs when the error terms across DMUs are correlated [21, 24, 47]. The main reason for contemporaneous correlation in DEA is that each DMU's score is influenced by the performance of efficient DMUs. If certain efficient DMUs systematically influence certain other DMUs, it may cause correlation among their error terms. Two tests for contemporaneous correlation when the number of panel members exceeds the number of time periods are Frees' R^2_{AVE} evaluated with his Q-distribution [20, 21], and Pesaran's CD cross-sectional dependence test [18]. When the number of periods exceeds the number of panel members, then the Breusch-Pagan Lagrange Multiplier test for contemporaneous correlation also can be applied [47].

GLS models that can simultaneously correct for serial and contemporaneous correlation, and for heteroskedasticity, are readily available. For example, when the number of DMUs exceeds the number of time periods, one can apply the Driscoll and Kraay [19] standard error estimator (available in Stata with

the PDA command ***xtscc***) [10]. This yields a nonparametric variance-covariance matrix estimator with standard errors that are robust to heteroskedasticity and to serial and contemporaneous correlation[27]. When the number of time periods exceeds the number of DMUs and a fixed effects model is used, then the preceding estimator as well as the Prais-Winsten estimator [24] can be utilized (available in Stata with the PDA command ***xtpcse***) [10, 42]. It computes parametric variance-covariance estimates that are robust to contemporaneous and serial correlation, and heteroskedasticity.

7. Pharmacy application

7.1 Data, inputs, and outputs

The dataset comes from a system of 12 hospitals in the US, and consists of 13 periods of bi-weekly data from the first six months of 2008. The data are from the pharmacy departments within these hospitals. This group of hospital pharmacies recently adopted a new set of clinical and distributional output indicators. In another paper that analyzes all of the inputs and outputs, we used a intertemporal frontier, so all 10 outputs and 3 inputs could be included and all efficiency scores would be estimated from the same efficient set [9]. For this paper, we use a contemporaneous frontier so the statistical diagnostic tests would be more sensitive to violations of independence. However, this results in only 12 observations for each DEA, so the number of inputs and outputs had to be decreased. Therefore, we use clinical outputs and input to measure the cross-sectional clinical efficiency of each pharmacy for 13 periods.

There is one input– clinical labor hours. Labor is the most important and controllable input that impacts hospital pharmacy efficiency, particularly from the standpoint of clinical pharmacy services. While the cost of drugs (another potential input) used in the production process in hospital pharmacy is a significant resource, it relates more to the product-related or "distributive" functions of the pharmacy and is primarily considered a pass-though (or "throughput"). All other operating and capital costs are relatively insignificant, are directly related to labor hours, and certainly could not be substituted for labor.

There are two outputs included in the analyses. These are the number of clinical interventions made by pharmacists, and the sum of the estimated dollar savings from those interventions. Clinical functions of hospital pharmacists include activities such as medication management, drug evaluation and selection, and

reviewing patient drug use. By making recommendations to physicians about the appropriateness of drug therapy, pharmacist time spent in clinical activities often results in substantial savings to the hospital in drug costs as well as improvements in patient outcomes. The number of interventions serves as a proxy for the positive effects on patients, and the dollar savings stands for itself. The data for each of the clinical outputs and inputs were available for each hospital and each pay period, and were generated by the pharmacy electronic documentation system used at the facilities.

7.2 Statistical panel data model

Our statistical model (Equation 2) is:

$$w_j\theta_{jt} = \alpha_j + \beta_j(t-1) + u_{jt} \qquad (2)$$

θ_{jt} is the efficiency score of pharmacy *j* in period *t*. The response variable is the product of θ_{jt} and the weight w_j, using Equation 3 (below) to compute the weight based on the standard error of estimate for data from pharmacy *j*. α_j is the individual effect of pharmacy *j*, β_j is the mean change per period in the individual effect α_j of pharmacy *j*, and u_{jt} is the random error in the response variable of pharmacy *j* in period *t*. Some pharmacies showed linear trends in efficiency over time, so Equation 2 includes a factor (*t*-1) that adjusts expected efficiency for the year involved. If there is no change in the efficiency scores over time for pharmacy *j*, or if the temporal trend is inconsistent, then β_j will not be significant. Finally, in Equation 3, $\alpha_j^* and \beta_j^*$ are estimated using θ_{jt} as the response variable. Because the pharmacies' error distributions are heteroskedastic, we obtain homoskedasticity by weighting each pharmacy *j*'s DEA scores by its standard error [11], per Equation 3.

$$w_j = \left(\sum_{t=1}^{T} (\theta_{jt} - \alpha_j^* - \beta_j^*(t-1))_{jt}^2 / T \right)^{-1/2} \qquad (3)$$

7.3 Results of Statistical Diagnostic Tests

The Breusch-Pagan/Cook-Weisberg test for heteroskedasticity found no statistically significant differences among the pharmacies [$\chi^2 = 0.00$, P($\chi^2(11) > 0.005) = 1.0000$]. The Shapiro-Wilk W test did not reject Normality [z =1.639, P(z > 1.639) = 0.051], nor did the Shapiro-Francia W' test [z =1.637, P(z > 1.637) = 0.051], nor did the skewness/kurtosis test for Normality [$\chi^2 = 3.89$, P($\chi^2(2) > 3.89) = 0.143$]. There was no statistically significant first-order serial correlation based on the Wooldridge test for autocorrelation in panel data [F = 0.494, P(F(1,11) > 0.494 = 0.497]. Tests for cross-sectional independence of residuals include Frees' R^2_{AVE} [R^2_{AVE} = -0.167, P(R^2_{AVE} > 0.1984 = 0.10], Pesaran's CD cross-sectional dependence test [CD = -2.039, P(CD > -2.039) = 0.958], and, since the number of time periods (13) slightly exceeds the number of pharmacies (12), the Breusch-Pagan LM test of independence is available [$\chi^2 = 66.667$, P($\chi^2(66) > 66.667) = 0.454$]. Therefore, the null hypotheses that the residuals are i.i.d. and Normally distributed cannot be rejected, so we employ the standard assumptions in developing probabilities and confidence intervals. The fixed-effect panel regression based on Equation 2 had an R-square of 0.565 [F = 7.47, P(F(23,132) >7.47) < 0.00005].

7.4 Confidence intervals

Figure 1 provides the range within which true efficiency (the actual value of the population mean) of each hospital pharmacy's clinical system is expected to occur with 0.95 confidence given the data. They are calculated as the expected value of the observation plus/minus the standard error of prediction times 1.96.

One point of interest is the relatively wide range within which the true efficiencies are expected to occur. For example, one would not be able to statistically reject at a 5% level, a hypothesis that the true efficiency was .7, or .8 for any of the hospitals. Though one could reject, with the exception of hospital pharmacy number 9, that true efficiency was below .6.

--put figure 1 about here—

7.5 Control charts

Figure 2 shows control charts for each pharmacy. If the most recent observation is at or below the bottom line of the chart, then this low efficiency would occur by chance only 10 percent of the time. Specifically, the bottom line is the expected value of the observation minus the standard error of forecast, that is, the standard error of the point prediction for one observation, times 1.282.

If the most recent observation is at or below the bottom line, it is likely that its efficiency score is not just a random variation, but really is lower than expected. In such cases, an immediate examination is in order. Such an examination is needed for hospital pharmacy 9, whose most recent efficiency level is very likely a reflection of true inefficiency rather than random variation.

--put figure 2 about here--

Because we had 12 hospitals but only 13 periods, we included all periods in our computations. However, in order to increase the power of the models to identify variations that are not random but a true change, the final period should be excluded from both the DEA and PDA models when developing control charts so its variation does not influence the results.

8. Transit application

8.1 Data, inputs and outputs

Urban transit agencies often oversee multiple modes and provider types of public transportation in their metropolitan areas. In the United States, the most common non-rail modes are scheduled motorbus, and paratransit (demand-responsive transit), with service being provided both by the agency itself and by outsourcing service. That is, U.S. transit agencies provide non-rail service with from one to four organizational subunits: directly-operated motorbus service, outsourced motorbus service, directly-operated demand-responsive service, and/or outsourced demand-responsive service. The largest agencies generally have two or three subunits, with all four subunits being utilized by some. [46]

We consider one output and one input from each subunit. We use the annual number of vehicle miles supplied by each subunit as our indicator of output, and each subunit's operating expenses (standardized for cost differences across time and across cities) as our indicator of input. Thus, there are four outputs

and four inputs. Our sample consists of the 50 agencies with 150 or more vehicles in maximum service, which included all such agencies for which all of the needed annual data were available for the years 2002-2006. The data are from the National Transit Database [46]. More detail on inputs, outputs and organization can be found in a paper providing a protocol for analyzing the efficiency of an urban area's transit when multiple types of service are operated [8].

8.2 Statistical panel data model

The PDA regression model (Equation 4) is

$$w_j \theta_{jt} = \alpha_j + u_{jt} \quad (4)$$

The definitions are the same as for Equation 2. However, unlike the model we used for the pharmacy data, we did not include a term for trends in the transit model. We did this partly because 5 observations seem too few to validly identify trends, and partly to demonstrate a model that estimates each DMU's true efficiency based on multiple observations. Equation 5 provides the model used to determine weights, [11] with definitions being the same as those for Equation 3.

$$w_j = \left(\sum_{t=1}^{T} (\theta_{jt} - \alpha_j^*)^2 / T \right)^{-3/2} \quad (5)$$

8.3 Results of Statistical Diagnostic Tests

The Breusch-Pagan/Cook-Weisberg test for heteroskedasticity found no statistically significant differences [$\chi^2 = 40.18$, P($\chi^2(49) > 40.18$) = 0.8113]. The Shapiro-Wilk W test did not reject Normality [z = 0.025; P(z > 0.025) = 0.49], nor did the Shapiro-Francia W' test [z = -0.532; P(z > -0.532) = 0.70], nor did the joint skewness/kurtosis test for Normality [$\chi^2 = 1.92$; $P(\chi^2(2) > 1.92) = 0.3829$]. There was statistically significant first-order serial correlation based on the Wooldridge test for autocorrelation in panel data [F = 19.981, P(F(1,49) > 19.981 < 0.0005]. Because the number of DMUs exceeds the number of time periods, the tests used for cross-sectional independence of residuals were Frees' R^2_{AVE} and Pesaran's CD cross-sectional dependence test. The results are R^2_{AVE} = 3.809, P(R^2_{AVE} > 1.1046) =0.01; and CD =

3.937, P(CD > 3.937) = 0.0001. Because both serial and contemporaneous correlations are present to a statistically significant degree, we must adjust for this in our subsequent confidence limit estimations.

8.4 Ranges within which true mean efficiencies occur

Based on the preceding diagnostics and the fact that the number of DMUs exceeds the number of time periods, we estimated the efficiencies of each DMU using the Driscoll-Kraay model, thereby correcting for serial correlation and for contemporaneous correlation. It is worth noting that this fixed-effect panel regression based on Equation 4 had an R-square of 0.995 [F = 835.55, P(F(49, 200) >835.55) < 0.00005].

The resulting efficiency estimates for the individual DMUs are shown in Table 1. The DEA model is input oriented, so lower scores mean lower efficiency. The confidence intervals are based on the standard error of prediction of the true expected value, at the 0.98 level of confidence. This means that any given system will only receive the incorrect efficiency report one percent of the time.

--put table 1 about here--

As shown in Table 1, the point estimates of the mean efficiency for 36 DMUs showed them to be inefficient, but 7 of them were not inefficient to a statistically significant degree. Of the 14 DMUs with efficient mean point estimates, 5 were not efficient to a statistically significant degree. Thus, whether a quarter of the agencies were or were not efficient cannot be determined with statistical confidence. Therefore, classifying a DMU as efficient (inefficient) based on a single DEA score or even a five-year mean, without considering confidence intervals, renders the validity of the classifications questionable.

9. Discussion

It may be worth noting that contemporaneous correlation does <u>not</u> bias the expected value of an estimated DEA score, but variance estimates can be more efficient if this correlation is taken into account. The main benefit of correcting for it when it exists is to attain better precision and thereby increasing power of the model to detect true differences. However, contemporaneous correlation should <u>not</u> automatically be corrected for unless it can be shown to be present. For example, using a GLS model that corrects for contemporaneous correlation when it is not in truth present will underestimate the model's confidence intervals, which could result in a DMU being incorrectly classified as efficient (inefficient).

We did not include any independent variables in our regressions, such as environmental or other exogenous influences on efficiency. They could be included if models correcting for any i.i.d. violations are used. But, we would counsel caution in any procedure that computes DEA scores in the first stage and then estimates the effect of exogenous variables in a later stage. As is well known [14, 25], such two-stage procedures can suffer from severe bias and precision problems, and sometimes lack sufficient power to detect the true effects of independent variables [4, 5, 48].

In order to improve a DMU's efficiency, it first is necessary to validly estimate the range within which its true efficiency occurs. Developing a methodology to do so is the purpose of this paper. After the efficiency indicators are collected, it next is necessary to identify the causes behind their values. However, illustrating this type of analysis is beyond the scope of this methodology-focused paper.

10. Conclusions

As exhibited in this paper, statistical panel data analysis methodology provides a useful tool for estimating valid confidence intervals for the DEA scores of individual organizations or other entities. PDA deals with stochastic data from both the individual entities and the production frontier, and greatly increases the variety and power of statistical models, diagnostic tests, and remedies. Moreover, although violations of i.i.d. and Normality by DEA score residuals can occur, they are <u>not</u> inevitable. In our pharmacy example, there were no violations except for heteroskedasticity. As we have demonstrated herein, PDA provides methodologies that can identify violations of i.i.d. and Normality when they do occur. And, PDA offers statistical models that can be used to remedy any violations, and thereafter estimate valid confidence intervals for individual DMUs.

Table 1. Efficiency Scores, 50 Large U.S. Transit Systems, 2002-2006

DMU_j	Superefficiency Scores, Input Oriented (0.98 Confidence Interval)			
	$E(\theta_j)$	Lower limit	Upper limit	Conclusion
1	0.779056	0.730532	0.82758	Inefficient*
3	1.021646	0.967599	1.075692	Efficient
8	0.858343	0.762322	0.954363	Inefficient*
1001	0.98948	0.93229	1.04667	Inefficient
1003	0.933634	0.849518	1.017749	Inefficient
1048	1.073378	1.039575	1.107182	Efficient*
2004	0.727619	0.705305	0.749934	Inefficient*
2007	0.812276	0.782443	0.842109	Inefficient*
2080	0.815464	0.754201	0.876727	Inefficient*
2113	0.814299	0.742971	0.885627	Inefficient*
3019	0.749426	0.628702	0.87015	Inefficient*
3030	0.901075	0.848103	0.954047	Inefficient*
3034	0.765765	0.680337	0.851193	Inefficient*
4003	1.300698	1.126881	1.474515	Efficient*
4008	1.38973	1.168751	1.610708	Efficient*
4018	0.954312	0.918467	0.990157	Inefficient*
4022	0.839205	0.694752	0.983658	Inefficient*
4029	1.698039	1.6242	1.771878	Efficient*
4035	1.933436	1.84864	2.018232	Efficient*
4041	0.768365	0.663284	0.873447	Inefficient*
4086	0.51111	0.46258	0.55964	Inefficient*
5005	1.029715	0.957623	1.101806	Efficient
5008	0.888916	0.823575	0.954258	Inefficient*
5012	0.963714	0.945931	0.981497	Inefficient*
5015	0.749967	0.720762	0.779173	Inefficient*
5016	1.096869	1.06981	1.123927	Efficient*
5022	1.013016	0.957384	1.068649	Efficient
5027	0.932835	0.903686	0.961984	Inefficient*
5031	0.903757	0.878678	0.928836	Inefficient*
5032	4.828623	4.091832	5.565414	Efficient*
5066	0.604844	0.587454	0.622233	Inefficient*
5113	0.947525	0.88366	1.011391	Inefficient
5119	0.546276	0.472372	0.620181	Inefficient*
6008	0.918045	0.850086	0.986004	Inefficient*
6011	1.040243	0.98791	1.092576	Efficient
6056	1.754943	0.498123	3.011763	Efficient
7005	1.264265	1.217288	1.311242	Efficient*
7006	0.73702	0.703953	0.770087	Inefficient*
8001	0.902934	0.774916	1.030952	Inefficient
8006	1.231883	1.155712	1.308054	Efficient*
9002	0.803977	0.756261	0.851693	Inefficient*
9009	0.915963	0.831508	1.000418	Inefficient
9013	0.697724	0.671099	0.724348	Inefficient*
9016	0.865849	0.724332	1.007365	Inefficient
9019	0.708942	0.681362	0.736521	Inefficient*
9023	0.794454	0.74184	0.847068	Inefficient*
9026	0.775653	0.723168	0.828139	Inefficient*
9030	0.876075	0.848637	0.903514	Inefficient*
9033	0.916297	0.822372	1.010222	Inefficient
9036	0.839873	0.748252	0.931494	Inefficient*

* Statistically significant at the 0.01 one-tailed level.

$E(\theta_j) < 1$ is not efficient

$E(\theta_j) \geq 1$ is efficient

Figure 1. 95% Confidence Interval for True Clinical System Efficiency, by Hospital Pharmacy

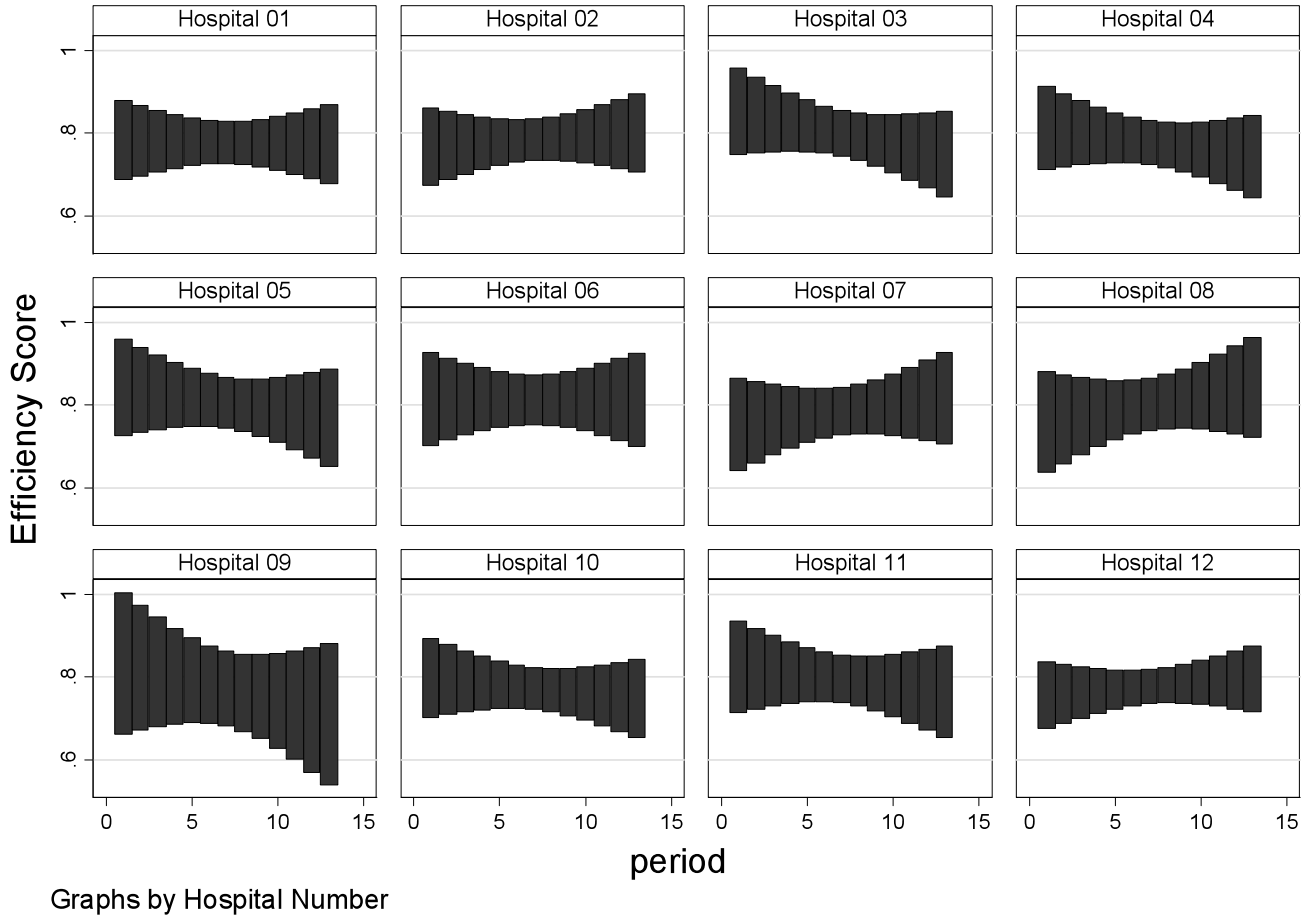
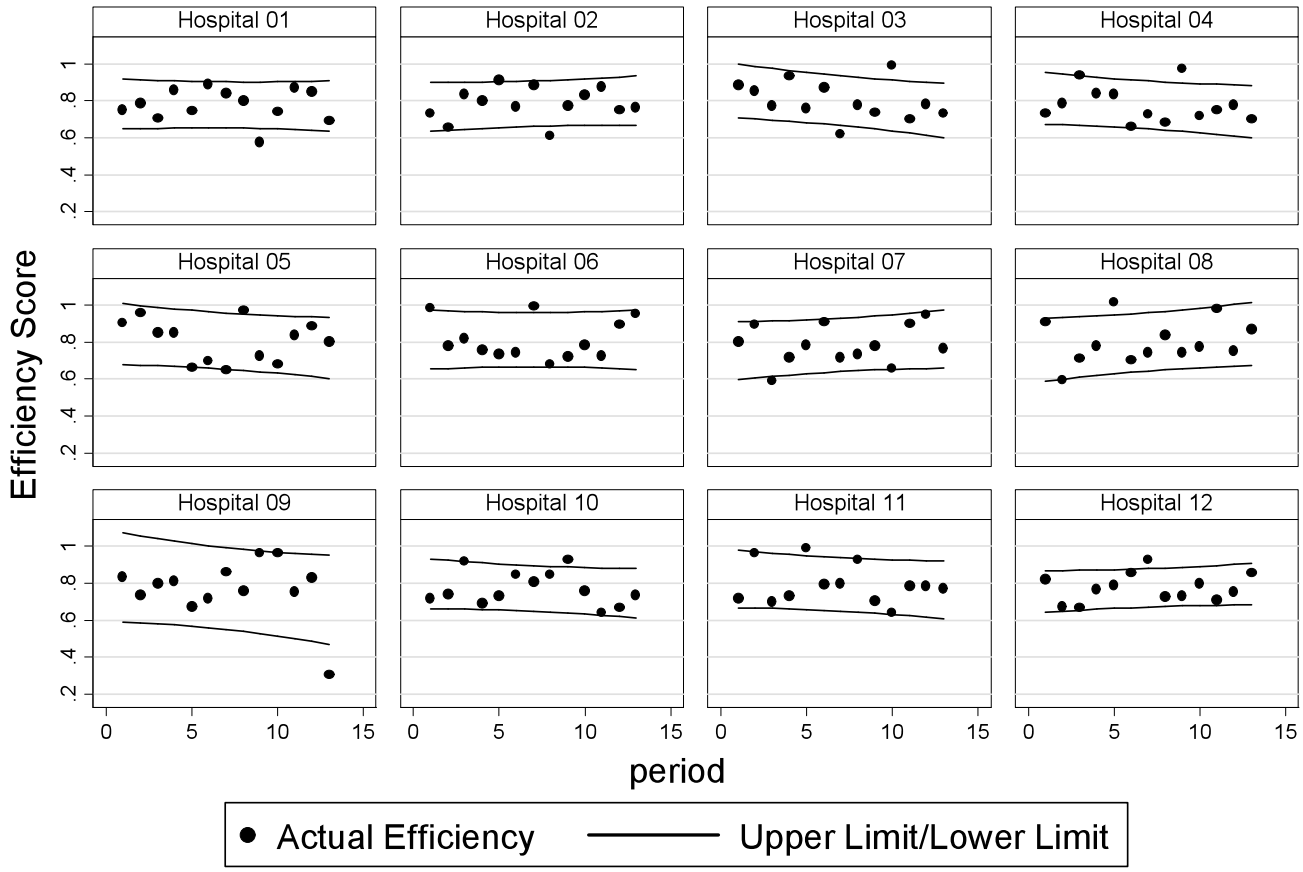


Figure 2. Control Charts by Hospital Pharmacy, and Reported Clinical Efficiencies. 10% probability that reported efficiency will be below bottom line by random chance



References

[1] S.E. Atkinson, and P.W. Wilson, *Comparing mean efficiency and productivity scores from small samples - a bootstrap methodology*, Journal of Productivity Analysis 6 (1995), pp. 137-152. Available at ISI:A1995RL08300003.

[2] B.H. Baltagi, *Econometric analysis of panel data,* ed, Vol. 4, John Wiley, West Sussex, England, 2008.

[3] R.D. Banker, and R. Natarajan, *Statistical tests based on DEA efficiency scores*, in *Handbook on data envelopment analysis*, W.W. Cooper, L.M. Seiford and J. Zhu eds., Kluwer, Boston, 2004, pp. 299-322.

[4] D.T. Barnum, and J.M. Gleason, *Bias and precision in the DEA two-stage method*, Applied Economics 40 (2008), pp. 2305-2311. Available at http://www.informaworld.com/10.1080/00036840600949470.

[5] D.T. Barnum, J.M. Gleason, and B. Hemily, *Using panel data analysis to estimate DEA confidence intervals adjusted for the environment*, Journal of Transportation Engineering 134 (2008), pp. 215-223.

[6] ---, *Using panel data analysis to estimate confidence intervals for the DEA efficiency of individual decision making units*, Applied Economics 41 (2009), pp. 3319-3326. Available at http://dx.doi.org/10.1080/00036840701493741.

[7] D.T. Barnum et al., *Progressing from uncertainty to risk for DEA-based decisions*, Journal of the Operational Research Society 61 (2010), pp. 1548-1555. Available at doi:10.1057/jors.2009.120.

[8] D.T. Barnum, M.G. Karlaftis, and S. Tandon, *Improving the efficiency of metropolitan area transit by joint DEA of its multiple providers*, Transportation Research Part E: Logistics and Transportation Review, In Press (2011). Available at http://ssrn.com/abstract=1399091.

[9] D.T. Barnum et al., *Improving the efficiency of distributive and clinical services in hospital pharmacy with DEA*, Journal of Medical Systems 35 (2011), pp. 59-70. Available at http://dx.doi.org/10.1007/s10916-009-9341-2.

[10] A.C. Cameron, and P.K. Trivendi, *Microeconometrics using Stata,* ed, Stata Press, College Station, Texas, 2009.

[11] R.J. Carroll, and D. Ruppert, *Transformation and weighting in regression*, *Monographs on statistics and applied probability,* ed, Chapman and Hall, New York, 1988.

[12] R.G. Chambers, and R. Färe, *Additive decomposition of profit efficiency*, Economics Letters 84 (2004), pp. 329-334. Available at http://www.sciencedirect.com/science/article/B6V84-4CF18G7-5/2/ba019d97c1de8260b86d2e1bbf930867.

[13] A. Charnes, W.W. Cooper, and E. Rhodes, *Measuring the efficiency of decision making units*, European Journal of Operational Research 2 (1978), pp. 429-444. Available at http://www.sciencedirect.com/science/article/B6VCT-48NBHPY-2BX/2/a0874a3b5efce500292fee5af445407c.

[14] T.J. Coelli et al., *An introduction to efficiency and productivity analysis,* ed, Vol. 2, Springer, New York, 2005.

[15] W.W. Cooper et al., *Chance constrained programming approaches to technical efficiencies and inefficiencies in stochastic data envelopment analysis*, Journal of the Operational Research Society 53 (2002), pp. 1347-1356. Available at http://proquest.umi.com/pqdweb?did=258937841&Fmt=7&clientId=8224&RQT=309&VName=PQD.

[16] ---, *Sensitivity analysis in DEA*, in *Handbook on data envelopment analysis*, W.W. Cooper, L.M. Seiford and J. Zhu eds., Kluwer Academic Publishers, Boston, 2004, pp. 75-97.

[17] W.W. Cooper, L.M. Seiford, and J. Zhu, *Data envelopment analysis: History, models and interpretations*, in *Handbook on data envelopment analysis*, W.W. Cooper, L.M. Seiford and J. Zhu eds., Kluwer Academic Publishers, Boston, 2004, pp. 1-40.

[18] R.E. De Hoyos, and V. Sarafidis, *Testing for cross-sectional dependence in panel-data models*, Stata Journal 6 (2006), pp. 482-496.

[19] J.C. Driscoll, and A.C. Kraay, *Consistent covariance matrix estimation with spatially dependent panel data*, Review of Economics and Statistics 80 (1998), pp. 549-560. Available at http://search.ebscohost.com/login.aspx?direct=true&db=bsh&AN=1450091&site=ehost-live.

[20] E.W. Frees, *Assessing cross-sectional correlation in panel data*, Journal of Econometrics 69 (1995), pp. 393-414.

[21] ---, *Longitudinal and panel data: Analysis and applications in the social sciences,* ed, Cambridge University Press, Cambridge, U.K, 2004.

[22] B.J. Gajewski et al., *On estimating the distribution of data envelopment analysis efficiency scores: An application to nursing homes' care planning process*, Journal of Applied Statistics 36 (2009), pp. 933-944. Available at http://search.ebscohost.com/login.aspx?direct=true&db=bsh&AN=44252744&site=ehost-live.

[23] J.M. Gleason, *Empirical tests of the intrinsic pseudorandom number generator on IBM-compatible microcomputers*, Computer Methods and Programs in Biomedicine 33 (1990), pp. 171-174. Available at http://www.sciencedirect.com/science/article/B6T5J-48V27B1-20/1/3093127028002962815d772c18d4b591.

[24] W.H. Greene, *Econometric analysis, 6th ed,* ed, Vol. 5, Prentice Hall, Upper Saddle River, NJ, 2008.

[25] S. Grosskopf, *Statistical inference and nonparametric efficiency: A selective survey*, Journal of Productivity Analysis 7 (1996), pp. 161-176. Available at ISI:A1996VB11200004.

[26] S.R. Gupta et al., *Association between hospital size and pharmacy department productivity*, American Journal of Health-System Pharmacy 64 (2007), pp. 937-944. Available at Publisher URL: www.cinahl.com/cgi-bin/refsvc?jid=1279&accno=2009573678;http://search.ebscohost.com/login.aspx?direct=true&db=rzh&AN=2009573678&site=ehost-live.

[27] D. Hoechle, *Robust standard errors for panel regressions with cross-sectional dependence*, Stata Journal 7 (2007), pp. 281-312.

[28] A. Hoff, *Bootstrapping Malmquist indices for Danish seiners in the North Sea and Skagerrak*, Journal of Applied Statistics 33 (2006), pp. 891-907. Available at http://search.ebscohost.com/login.aspx?direct=true&db=bsh&AN=23253113&site=ehost-live.

[29] A. Kittelson et al., *A guidebook for developing a transit performance-measurement system, Transit Cooperative Research Program Report 88*, Transportation Research Board, Washington, D.C., 2003.

[30] S. Kumbhakar, and C.A.K. Lovell, *Stochastic frontier analysis,* ed, Cambridge University Press, Cambridge England, 2000.

[31] J. Lin, P. Wang, and D.T. Barnum, *A quality control framework for bus schedule reliability*, Transportation Research Part E: Logistics and Transportation Review 44 (2008), pp. 1086-1098. Available at http://www.sciencedirect.com/science/article/B6VHF-4R8PNVM-1/2/9a717594cbe0b84661573f343ffe5806.

[32] V. Perk, and N. Kamp, *Benchmark rankings for transit systems in the United States*, National Center for Transit Research, Univ of South Florida, 2004.

[33] H. Scheel, *EMS: Efficiency Measurement System* 2008.

[34] G.T. Schumock et al., *Data envelopment analysis - A method for comparing hospital pharmacy productivity*, American Journal of Health-System Pharmacy (2009).

[35] L. Simar, and P.W. Wilson, *Sensitivity analysis of efficiency scores: How to bootstrap in nonparametric frontier models*, Management Science 44 (1998), pp. 49-61. Available at ISI:000072627200004.

[36] ---, *Estimating and bootstrapping Malmquist indices*, European Journal of Operational Research 115 (1999), pp. 459-471. Available at http://www.sciencedirect.com/science/article/B6VCT-3W19817-4/2/e122e01034931278e89be4818ba4aefa.

[37] ---, *A general methodology for bootstrapping in non-parametric frontier models*, Journal of Applied Statistics 27 (2000), pp. 779-802. Available at ISI:000088600800011.

[38] ---, *Statistical inference in nonparametric frontier models: The state of the art*, Journal of Productivity Analysis 13 (2000), pp. 49-78. Available at ISI:000084500700003.
[39] ---, *Estimation and inference in two-stage, semi-parametric models of production processes*, Journal of Econometrics 136 (2007), pp. 31-64. Available at http://www.sciencedirect.com/science/article/B6VC0-4H2PJVP-1/2/ccb96a8862f61854e86fd3cb4d014e5b.
[40] R.G. Stanley, and P.G. Hendren, *Performance-based measures in transit fund allocation: A synthesis of transit practice. Transit Cooperative Research Program Synthesis 56*, Transportation Research Board, 2004.
[41] StataCorp, *Stata Base Reference Manual Release 10,* ed, StataCorp, College Station, Texas, 2007.
[42] ---, *Stata statistical software release 10: Longitudinal/panel data [XT],* ed, StataCorp, College Station, TX, 2007.
[43] ---, Stata/IC 10 for Windows; software available at www.stata.com.
[44] C. Tser-yieth, *A comparison of chance-constrained DEA and stochastic frontier analysis: Bank efficiency in Taiwan*, Journal of the Operational Research Society 53 (2002), pp. 492-500. Available at http://proquest.umi.com/pqdweb?did=119356958&Fmt=7&clientId=8224&RQT=309&VName=PQD.
[45] H. Tulkens, and P. Vanden Eeckaut, *Nonparametric efficiency, progress and regress measures for panel-data - methodological aspects*, European Journal of Operational Research 80 (1995), pp. 474-499. Available at ISI:A1995QC23600004.
[46] United_States_Federal_Transit_Administration, *National Transit Database*, United States Federal Transit Administration,, 1997- 2008.
[47] J.M. Wooldridge, *Econometric analysis of cross section and panel data,* ed, MIT Press, Cambridge, MA, 2002.
[48] V. Zelenyuk, *Power of significance test of dummies in Simar-Wilson two-stage efficiency analysis model*, Institut de Statistique, UniversitÇ Catholique de Louvain, Belgium, 2005.