Unobserved Heterogeneity and Efficiency Measurement in Public Transport

by

Astrid Cullmann, Mehdi Farsi and Massimo Filippini

Address for correspondence:
Astrid Cullmann, DIW Berlin (German Institute for Economic Research), Department of Innovation, Manufacturing, Service, Mohrenstrasse 58, 10117 Berlin, Germany (acullmann@diw.de).
Mehdi Farsi is at the Faculty of Economics, University of Neuchâtel;
Massimo Filippini is at D-MTEC, ETH Zurich and also at the Department of Economics, University of Lugano.

ABSTRACT

Efficiency measurement in public transport requires an adequate account of unobserved network characteristics that are typically modeled as factors separable from the production process. This paper proposes a panel data model that allows for non-separable firm-specific heterogeneity in an input distance function. The proposed model is applied to a sample of German and Swiss urban transit companies operating from 1991 to 2006. The results underline the presence of non-separable unobserved factors and their effects on technological characteristics such as returns to scale. Moreover, the data suggest that the effect of time-invariant heterogeneity could be significantly greater than technical inefficiency.

Date of final version: 27th July 2010

1.0 Introduction

Following the explosive growth of subsidy requirements for public transport services in the 1970s and 1980s, several European governments have gradually introduced regulatory reforms in their local transport sectors. Most of these countries, in line with the EU directives, have adopted a competitive tendering procedure for the assignment of franchised monopolies to local service providers. Competitive tendering is expected to induce relatively strong incentives for cost efficiency. However, as documented in several studies (Toner, 2001; Boitani and Cambini, 2002; Cambini and Filippini, 2003) these procedures have experienced many implementation obstacles resulting in a tendency toward auctioning small networks with suboptimal scale and density as well as potential collusion among the bidders. An alternative approach would be incentive regulation schemes, such as yardstick competition or performance based contracts.[1] These schemes are based on benchmarking analysis of costs and/or quality to determine the transfers and prices.

In Switzerland and Germany competitive tendering has been introduced but remains limited to certain areas.[2] Nevertheless, regional authorities have been discussing the possibility of adopting high-powered contracts based on yardstick competition as in Shleifer (1985). In this context benchmarking namely, estimating companies' productive efficiency could be used as a complementary control instrument in determining subsidies and prices.[3] However, given the observed sensitivity of benchmarking methods,[4] the

[1] For a general discussion on these two approaches see Demsetz (1968), Laffont and Tirole (1993), Klemperer (1999), Hensher (2007) and Hensher and Stanley (2003). In particular, the latter two studies have shown that performance based contracts can reach a greater social surplus than competitive tendering.
[2] These include Swiss rural areas, one German state (Hesse) and only a few large German cities (Hamburg and Munich). In most other cases, particularly, in Swiss urban areas, concessions are granted to incumbent providers without any risk of competitive tendering.
[3] For an application of yardstick competition in the transport sector see Dalen and Gòmez-Lobo (2003).

reliability of efficiency estimates depends on an adequate modeling of firms' unobserved heterogeneity.

Since urban transit companies operate in different networks and environments, and provide urban passenger services using a diversity of vehicles (bus, tramway, light rail, etc.) there are a great number of factors that affect the production process. Benchmarking methods have been subject to a strong criticism, mainly because many of these firm-level differences are not usually observed by the analyst. Moreover, certain characteristics such as network shape and complexity remain omitted from the models because they are not easily measurable by single factors amenable to benchmarking techniques. Therefore, unobserved firm heterogeneity is inevitably an important part of measuring efficiency in public transport.

Thus, our main objective is to derive and apply an appropriate Stochastic Frontier (SF) model, which is able to capture firm-specific unobserved heterogeneity using panel data. In recent SF panel data models such as Greene (2004, 2005a,b) unobserved firm-specific heterogeneity is mainly modeled as an additive stochastic factor represented by conventional fixed or random effects. Within this framework the unobserved factors are considered as separable factors from the production process. In this paper, we argue that the entire production process is organized around the network structure. In line with Bagdadioglu and Weyman-Jones, (2008) we assume that the unobserved heterogeneity is inevitably non-separable from the production process thus interrelated with the observed input and output factors.

[4] See Jamasb and Pollit (2003), Estache et al. (2004) and Farsi et al. (2006b) for examples.

From a methodological point of view, the analysis contributes to the discussion of unobserved heterogeneity that is particularly relevant for efficiency measurement in network industries. The proposed method has been applied to a sample of German and Swiss public transport companies. The results indicate that the unobserved heterogeneity could dominate the efficiency differences. Consistent with previous studies these results point to the importance of the underlying assumptions used to distinguish between inefficiency and unobserved firm differences. The rest of the paper is organized as follows: Sections 2 presents the model specification. The data and the econometric models are explained in Section 3. Section 4 presents the estimation results and discusses their implications, and Section 5 provides the conclusions.

2.0 Model Specification

There is a great body of literature on the estimation of production and cost frontiers for public transit operators.[5] However, the majority of these studies estimate single output production or cost frontiers. There are only a few studies that estimated a multi-output cost function. The most relevant ones in this category are Viton (1992), Viton (1993) and Colburn and Talley (1992), both of which analyzed the long run cost structure of urban multi-mode transit system in the U.S. Viton (1992) studied the cost structure of a sample of 289 urban transit companies operating in the U.S. between 1984 and 1986. Six modes are distinguished: motor-bus, rapid-rail, streetcar, trolley-bus, demand responsive mode and a last mode including all other modes. Viton uses a quadratic total cost function. Colburn and Talley (1992) analyze the economies of scale and scope of a single urban multi-service company using quarterly data from 1979 to 1988. Four modes are

[5] See De Borger et al. (2002) for a detailed literature review.

distinguished: motor-bus, dial-a-ride, elderly service, and van pool service. Colburn and Talley used a translog total cost function. The first European analysis for multi-output firms has been performed by Farsi et al. (2006b). In this study, the authors estimate a quadratic cost function considering three modes (motor-bus, streetcar, trolley-bus) and using a dataset composed of 16 Swiss multi-mode urban transport operators observed during the period 1985–2003. None of these studies estimated a frontier function and, therefore, did not perform an efficiency analysis. The main interest of these studies was in the estimation of the economies of scale and scope.

To measure the efficiency level of the multi-outputs Swiss and German urban transit companies we apply a parametric frontier input distance function.[6] We therefore focus on the technical inefficiency as opposed to possible inefficiencies due to suboptimal allocation of input factors. Because of the lack of consistent data on costs and input prices especially in the case of Germany, we could not use a multi-output cost function. Compared to production functions the distance functions are more readily adaptable to multi-output contexts. In addition, the choice of distance functions does not require the cost minimization assumption.[7] One concern in the econometric estimation might be the regressor endogeneity which may introduce possible simultaneous equation bias.[8] Sickles et al. (2002) and Atkinson and Primont (2002) used methods based on instrumental variables to correct for such endogeneities. However, Coelli (2002) showed that compared to production functions, the distance functions do not face a greater risk of

[6] For the use of parametric distance functions in the transport sector see Coelli and Perelman (1999, 2000).
[7] For a discussion on the advantages and drawbacks of the distance-functions approach see Coelli (2002) and Coelli and Perelman (2000).
[8] This results from the fact that for instance in an input distance function, the inputs appearing on the right hand side of the equation might be correlated with the residuals.

endogeneity bias.[9] Assuming that outputs are exogenous for given companies, we favored an input distance specification as opposed to an output distance function.[10]

The input distance function is defined on the input set as the extent to which the input vector may be proportionally contracted with the output vector held fixed (see Coelli, 2002):

$$d_I(x,y) = \max\{\rho : (x/\rho) \in L(y)\} \quad (1).$$

$d_I(x,y)$ will take a value greater than or equal to one if the input vector x is an element of the feasible input set $L(y)$. In addition, $d_I(x,y)=1$ if x is located on the inner boundary of the input set. ρ represents the scalar distance, so the amount by which the input vector can be deflated. It is assumed that the technology satisfies the standard axioms: $d_I(x,y)$ is non-decreasing, linearly homogeneous and concave in x and decreasing in y.[11]

[9] A second issue is that estimated input distance functions often fail to satisfy the concavity properties implied by economic theory. Regularity conditions could also be imposed by estimating the model in a Bayesian framework (see O'Donnell and Coelli, 2005).

[10] An input-oriented distance function is motivated by the nature of production in the public transport sector, because it implies that efficiency is improved by reducing input usage for a given exogenous output, set by regulators or the demand side factors that are beyond the provider's control.

[11] See Coelli (2002) and Färe and Primont (1995) for more details on these properties.

For the specification of the model we considered public transit companies characterized by a production process with three inputs and two outputs. Following Farsi et al. (2006a, 2006b) we consider two purely supply-oriented measures of the output namely, seat-kilometers provided by tramways and buses respectively.[12] Labor input, number of trams and number of buses are considered as input factors. The input distance function can be accordingly specified as:

$$d = f(X_L, X_{CT}, X_{CB}, Y_T, Y_B, Z, \gamma, t) \qquad (2),$$

where x_L is labor input and x_{CB}, x_{CT} are respectively two indicators of the capital input, number of buses and number of tramways. y_B and y_T are the numbers of seat-kilometers provided by buses and tramways respectively. t is a time variable which captures the shift in technology, Z is the total network length (trams and bus networks) introduced in the model in order to capture part of the observable heterogeneity of the operating environment of the companies, and γ is a time-invariant stochastic term that represents all the unobserved structural characteristics of the network.

As in most empirical studies in the production literature, we specify a translog functional form in order to satisfy flexibility while allowing a straightforward imposition of linear homogeneity.[13] The adopted model in (2) might appear a rather parsimonious

12 We concentrate our analysis only on transit companies supplying services using the same transport modes (buses and tramways). Therefore, we excluded transit companies operating with underground system as well as small companies that use only buses. Moreover, in Switzerland some of the companies supply trolley as well as autobus services. We assumed for the empirical analysis that the trolley busses feature similar characteristics as the autobuses, therefore we sum up both singles branches to have an aggregated bus stock and aggregated supplied services.

13 Following Lovell et al. (1994) and Coelli and Perelman (2000), a convenient method for imposing linear homogeneity constraint is to divide the inputs by one of the input factors. In translog form the input distance function is invariant to which input is chosen as the numéraire.

model that does not include some of the observed characteristics available in the data such as the size of service area and covered population as well as number of seats in each company's fleet. However, due to strong correlation of these variables with network length and other variables, models with additional variables face a great risk of multicollinearity that is particularly exacerbated because of the second-order terms in the translog form. Our preliminary analyses using several alternatives have favored the adopted specification above in terms of model's explanatory power as well as plausibility of the estimated coefficients.

Recognizing that the network length controls for only a part of network heterogeneity, we assume that the remaining factors that are constant over time, in particular those related to the shape and complexity[14] of the network are captured by the stochastic variable γ.

Assuming non-separability of the unobserved network structural variable, γ, the translog formulation of the model in equation (2) can be expressed as follows:

[14] Using a complexity indicator based on graph theory, Filippini and Maggi (1992) have shown the importance of network complexity in a cost function for transport companies. Unfortunately, we do not have data on the shape and structure of the networks.

$$
\begin{aligned}
\ln d_{it} = {} & \alpha_0 + \eta_1\gamma_i + \ln\left(x_{Lit}\right) + \alpha_{CT}\ln\frac{x_{CTit}}{x_{Lit}} + \eta_4\gamma_i\ln\frac{x_{CTit}}{x_{Lit}} + \alpha_{CB}\ln\frac{x_{CBit}}{x_{Lit}} \\
& + \eta_3\gamma_i\ln\frac{x_{CBit}}{x_{Lit}} + \frac{1}{2}\alpha_{CTCT}(\ln\frac{x_{CTit}}{x_{Lit}})^2 + \frac{1}{2}\alpha_{CBCB}(\ln\frac{x_{CBit}}{x_{Lit}})^2 + \frac{1}{2}\eta_2\gamma_i^2 \\
& + \alpha_{CBCT}(\ln\frac{x_{CBit}}{x_{Lit}})(\ln\frac{x_{CTit}}{x_{Lit}}) + \beta_T\ln y_{Tit} + \eta_6\gamma_i\ln y_{Tit} + \beta_B\ln y_{Bit} \\
& + \eta_5\gamma_i\ln y_{Bit} + \frac{1}{2}\beta_{TT}(\ln y_{Tit})^2 + \frac{1}{2}\beta_{BB}(\ln y_{Bit})^2 \\
& + \beta_{BT}(\ln y_{Bit})(\ln y_{Tit}) + \delta_{TCT}(\ln y_{Tit})(\ln\frac{x_{CTit}}{x_{Lit}}) + \delta_{BCT}(\ln y_{Bit})(\ln\frac{x_{CTit}}{x_{Lit}}) \\
& + \delta_{TCB}(\ln y_{Tit})(\ln\frac{x_{CBit}}{x_{Lit}}) + \delta_{BCB}(\ln y_{Bit})(\ln\frac{x_{CBit}}{x_{Lit}}) + \alpha_Z\ln Z_{it} + \eta_7\gamma_i\ln Z_{it} \\
& + \frac{1}{2}\alpha_{ZZ}(\ln Z_{it})^2 + \alpha_{ZB}(\ln y_{Bit})(\ln Z_{it}) + \alpha_{ZT}(\ln y_{Tit})(\ln Z_{it}) \\
& + \alpha_{ZCT}(\ln Z_{it})(\ln\frac{x_{CTit}}{x_{Lit}}) + \alpha_{ZCB}(\ln Z_{it})(\ln\frac{x_{CBit}}{x_{Lit}}) + \alpha_t t + v_{it} \qquad (3),
\end{aligned}
$$

where subscripts *i* and *t* denote the company and year respectively and v_{it} represents the additive residuals as a random error term. $\ln d_{it}$ is a nonnegative variable which can be associated with technical inefficiency u_{it}. A radial input-oriented measure of technical efficiency can be obtained by $TE = \frac{1}{d_{it}} = \exp(-u_{it})$. As we will see in the following section the model in (3) can be formulated as a common SF model with the combined error term $v_{it} - u_{it}$.

3.0 Data and econometric specification

3.1. Data

The sample used in this study is composed of an unbalanced panel data from Swiss and German transit companies that provide motor bus and tramway transport services. The data include 13 annual observations from 56 companies including 49 German and 7

Swiss companies. The sample period differs across the countries, covering from 1994 to 2006 for the case of Germany and 1991 to 2003 for the Swiss companies. In both cases the companies in the sample can be defined as independent local monopolies, given the fact that there is no overlap between the offered transport services across the companies.

The data for Germany is provided by the VDV Statistics.[15] Data are available for 360 public transport companies; among which 60 offer bus transport as well as regional rail services. We created a balanced panel data set for 49 multi-output companies offering tram and motor bus services in medium and large German cities.[16] In the case of Switzerland, all the local public transit services within the country's urban centers are covered by sixteen companies. For our analysis we selected seven Swiss companies that offer motor bus and tram transport.[17] For the years between 1991 and 1997 the Swiss data has been extracted from the annual statistics on public transport reported by the Swiss Federal Statistical Office (BFS (1991-97)). The data for the following years (1998-2003) have been collected from companies' annual reports. A descriptive summary of the data is given in Table 1.

Table 1

The companies included in the sample are characterized by a potentially strong heterogeneity in technologies, regulation restrictions, environmental variables and in

[15] VDV (*Verband Deutscher Verkehrsunternehmen*) or the Association of German Transport Companies represents about 440 member companies operating in public transport and freight.
[16] In order to have a more or less uniform sample we excluded four large companies (operating in Berlin, Hamburg, Munich and Nuremberg) that offer underground railway transport and three small single-output tolley-bus operators.
[17] We excluded the companies that offer trolley-bus services and those that are specialized in a single mode of transport.

particular network complexities. This large output heterogeneity is not completely observed in the data and evidently become more relevant for cross-country efficiency analyses. In the next section we describe how panel data models have been used in order to separate such unobserved factors from inefficiencies.

3.2 Econometric Specification using panel data

The first use of panel data models in stochastic frontier models goes back to Pitt and Lee (1981) who interpreted the panel data random effects as inefficiency rather than heterogeneity.[18] A main shortcoming of these models is that any unobserved, time-invariant, firm-specific heterogeneity is considered as inefficiency. In order to solve this problem, the SFA model in its original form (Aigner et al., 1977) can be readily extended to panel data models, by adding a fixed or random effect in the model. Although similar extensions have been proposed by several previous authors,[19] Greene (2005a,b) provides effective numerical solutions for both models with random and fixed effects, which he respectively refers to as "true" fixed and random effects models. Several recent studies such as Greene (2004), Farsi et al. (2006b), Alvarez et al. (2004) and Tsionas (2002) have followed this line. Some of these models have proved a certain success in a broad range of applications in network industries in that they give more plausible efficiency

[18] Pitt and Lee (1981)'s model is different from the conventional RE model in that the individual specific effects are assumed to follow a half-normal distribution. Important variations of this model were presented by Schmidt and Sickles (1984) who relaxed the distribution assumption and used the GLS estimator, and by Battese and Coelli (1988) who assumed a truncated normal distribution. In more recent papers the random effects model has been extended to include time-variant inefficiency. Cornwell et al. (1990) and Battese and Coelli (1992) are two important contributions in this regard. In particular the former paper proposes a flexible function of time with parameters varying among firms. However, in both these models the variation of efficiency with time is considered as a deterministic function that is commonly defined for all firms.

[19] In particular Kumbhakar (1991) and Heshmati and Kumbhakar (1994) proposed a three-stage estimation procedure to solve the model with time- and firm-specific effects.

estimates.[20] These results raise an important question as to what extent panel data models can be used for a better understanding of the inefficiencies and whether they can provide a reliable basis for benchmarking and incentive regulation systems in industries characterized by strong heterogeneity. This question is especially important when companies operate in multiple networks, entailing several network-specific heterogeneity dimensions. In most SF models the unobserved factors are widely modeled as separable factors from the production process (Greene, 2005a,b). However, we argue that the entire production process is organized around the network structures. Therefore, the unobserved heterogeneity is inevitably non-separable from the observed inputs and outputs. We propose a model assuming that unobserved heterogeneous factors are non-separable from the production process (see for instance Bagdadioglu and Weyman-Jones, 2008).

Along with the variation over time, the distinction between separable and non-separable factors can be helpful in disentangling the inefficiency from the unobserved firm-specific factors: Assuming that firm-specific factors are time-invariant but non-separable, while the inefficiency components are time-variant and separable, one can achieve a more realistic separation between the two components. In fact, being an integrated part of the technology process the unobserved network characteristics are non-separable but more or less time-invariant. Whereas it is likely that the main driving factor behind technical inefficiency namely, the management's efforts and incentives are independent from the production technology thus separable but, as shown by Alvarez et al. (2004), time-variant.

[20] See Saal et al. (2007), Farsi et al. (2005, 2006a,b) for applications in water distribution, electricity networks, bus transport and railroads respectively.

Considering the technical efficiency as a time-variant stochastic term with half-normal distribution, $u_{it} \sim N^{+}(0,\sigma_u^2)$, and an additive idiosyncratic symmetric error with normal distribution, $v_{it} \sim N(0,\sigma_v^2)$, the distance from the stochastic frontier ($\ln d_{it}$) can be specified as $v_{it} - u_{it}$. By substituting for $\ln d_{it}$ the stochastic frontier given in equation (3) can therefore be transformed to a random parameter stochastic frontier model with a single time-invariant random effect γ_i, as follows:

$$
\begin{aligned}
-\ln(x_{Lit}) = {} & \alpha_0 + \eta_1\gamma_i + \frac{1}{2}\eta_2\gamma_i^2 + (\alpha_{CT} + \eta_3\gamma_i)\ln\frac{x_{CTit}}{x_{Lit}} + (\alpha_{CB} + \eta_4\gamma_i)\ln\frac{x_{CBit}}{x_{Lit}} \\
& + \frac{1}{2}\alpha_{CTCT}(\ln\frac{x_{CTit}}{x_{Lit}})^2 + \frac{1}{2}\alpha_{CBCB}(\ln\frac{x_{CBit}}{x_{Lit}})^2 + \alpha_{CBCT}(\ln\frac{x_{CBit}}{x_{Lit}})(\ln\frac{x_{CTit}}{x_{Lit}}) \\
& + (\beta_T + \eta_5\gamma_i)\ln y_{Tit} + (\beta_B + \eta_6\gamma_i)\ln y_{Bit} + \frac{1}{2}\beta_{TT}(\ln y_{Tit})^2 \\
& + \frac{1}{2}\beta_{BB}(\ln y_{Bit})^2 + \beta_{BT}(\ln y_{Bit})(\ln y_{Tit}) + \delta_{TCT}(\ln y_{Tit})(\ln\frac{x_{CTit}}{x_{Lit}}) \\
& + \delta_{BCT}(\ln y_{Bit})(\ln\frac{x_{CTit}}{x_{Lit}}) + \delta_{TCB}(\ln y_{Tit})(\ln\frac{x_{CBit}}{x_{Lit}}) + \delta_{BCB}(\ln y_{Bit})(\ln\frac{x_{CBit}}{x_{Lit}}) \\
& + (\alpha_Z + \eta_7\gamma_i)\ln Z_{it} + \frac{1}{2}\alpha_{ZZ}(\ln Z_{it})^2 + \alpha_{ZB}(\ln y_{Bit})(\ln Z_{it}) + \alpha_{ZT}(\ln y_{Tit})(\ln Z_{it}) \\
& + \alpha_{ZCT}(\ln Z_{it})(\ln\frac{x_{CTit}}{x_{Lit}}) + \alpha_{ZCB}(\ln Z_{it})(\ln\frac{x_{CBit}}{x_{Lit}}) + \alpha_t t + v_{it} - u_{it}
\end{aligned}
\tag{4}
$$

We assume that the generic random effect γ_i follows a standard normal distribution, $N(0,1)$. With this assumption the econometric specification of the model is exactly similar to the 'fixed management model' proposed by Alvarez et al. (2004).[21] As

[21] It should be noted that Alvarez et al. (2004)'s interpretation of the latent variable in their model as a proxy for management's fixed input (effort), leading to an interrelation between inefficiency and the generic random effect , γ_i. Here, we assume that γ_i is an exogenous characteristic of the network thus independent of efficiency term, u_{it}.

it can be seen in equation (4), the latent variable γ_i enters in the model's intercept in a quadratic form as: $\alpha_0 + \eta_1\gamma_i + \frac{1}{2}\eta_2\gamma_i^2$, creating a skewed additive random effect, composed of a normal variable plus a Chi-squared with one degree of freedom (Greene, 2007). Moreover, the coefficients of all the first order terms of inputs $(\alpha_{CT} + \eta_3\gamma_i), (\alpha_{CB} + \eta_4\gamma_i)$, outputs $(\beta_T + \eta_5\gamma_i), (\beta_B + \eta_6\gamma_i)$, and the structural variable network length $(\alpha_Z + \eta_7\gamma_i)$ will become random coefficients with a common random effect, whereas all the coefficients of the second-order terms remain fixed. The random parameter model in (4) is estimated using the Simulated Maximum Likelihood module provided in *LIMDEP* 9.0.[22]

In summary, we see that the unobserved firm-specific heterogeneity attributed to the different network structures of the transport companies applies to marginal products represented by the coefficients of the distance function (see Section 4.1). We therefore allow firms to have different underlying production technologies caused by unobserved differences in technological conditions and network structures. In particular network structural characteristics play an important role in the production of transport services and cannot be fully captured by a production frontier with fixed coefficients. The proposed random coefficient frontier accounts for these differences using a single stochastic variable that is interpreted as an aggregate measure of unobservable structural characteristics that remain constant over time. We also use a special case of the model with complete separability, in which case, the random variable γ_i disappears from all the coefficients except the intercept.

[22] See Greene (2007) for more details on the numerical algorithm and choice of random draws.

4.0 Empirical results

Table 2 shows the regression results of the distance function, based on the stochastic frontier model given in equation 8. The table also includes the results of an alternative specification in which the unobserved network variable (γ_i) is assumed to be separable from all production factors. Given that all the variables are in logarithmic form, these coefficients can be directly interpreted as elasticities. For instance, the derivative of a translog input distance function with respect to a particular input is equal to the input contribution share of that input. In the interpretation of the coefficients it should be noted that a positive coefficient implies a contraction of the feasible input set thus, an increase in the distance function. Conversely, the negative effects are associated with an expansion in the input set. Therefore, outputs are expected to have negative coefficients while inputs are associated with positive effects. Similarly any positive coefficient indicates an improvement in production feasibilities, while negative coefficients can be interpreted as more resources and costs. For instance, the value of the coefficient of the time trend indicates an average technological progress of about 2 per cent per year over the sample period.

Table 2

The estimated coefficients (means for the random parameters) of the first-order terms have the expected signs and are statistically significant at the sample median. As expected, the coefficients of first-order output variables are negative and significantly

different from zero implying that the estimated distance function is decreasing in outputs. The coefficients of the first-order terms of the capital and labor inputs are as expected positive and significantly different from zero. The sum of the coefficients of the two output variables is 0.79 or 0.82 (depending on the model). This result suggests the presence of economies of density at the sample median, because, *ceteris paribus*, by increasing both outputs by 10 per cent, the input requirement will increase only by about 8 per cent. As for the effect of network length, the results show that the first order term is, as expected negative and statistically significant. The sum of this coefficient with the two coefficients of the two output variables is 0.87 or 0.82. This result indicates the presence of economies of scale, because by increasing both outputs and network length by 10 per cent, the input requirement will increase only by about 8 (9) per cent.[23]

The negative coefficients of the output square terms for both bus and tram outputs, suggest that the rate of economies of scale is decreasing in each output. The positive coefficient of the interaction of the two outputs indicates cost-complementarity between tram and bus services. For instance, the results suggest that increasing one output by 10 per cent, will result in 0.9 or 1.1 per cent (depending on the model) decrease in the marginal cost of the other output. The effect of interactions with the network length suggest that providing bus services over longer networks is relatively less costly, while for trams, longer networks are associated with higher marginal costs. This result is consistent with the fact that in tramways, the maintenance of the network infrastructure (rails and cables) in longer network might take relatively more capital and labor resources than in bus transport.

23 Note that in translog form, any statement about sample points other than the approximation point (here, sample median), should consider the second-order terms in addition to the main effects.

The table shows that in both models, the coefficients of the unobserved structural variable ($\eta_1 - \eta_7$) are significantly different from zero at conventional 5 per cent levels of significance. This provides empirical evidence for the presence of unobserved heterogeneity. Using a Wald test we tested the hypothesis of separability. The results (also listed in the table) favor the complete model, indicating that the unobserved network characteristics are not separable from observed production factors. Comparing the results across the two models indicates a close similarity in the coefficients of the first-order terms, suggesting that the estimates of returns to scale and other technological characteristics at the approximation point (here the sample median) are not sensitive to the assumption of separability. However, most second-order terms especially those related to network length (variable *Z*), vary across the two models. This suggests that quantities such as complementarity effects between different outputs as well as substitution elasticities between inputs could be sensitive to the assumptions related to separability from the unobserved network characteristics. The differences of second-order effects across the two models also suggest that the variation of the economies of scale at different levels of output and network length is sensitive to the separability assumption.

Studying the coefficients of the latent heterogeneity can be helpful in detecting the effects captured by that variable. The positive sign of the constant (η_1) indicates that higher levels of the latent variable (γ) are associated with network and environmental characteristics that are beneficial to production. Therefore the latent variable γ can be interpreted as an aggregate indicator of network structural characteristics with an inverse correlation with network complexity. With this interpretation in mind, namely associating lower values of γ with greater network complexity, we can explore the consistency of the

regression results with our underlying assumptions about network heterogeneity. The coefficients of the interactions of the unobserved heterogeneity with both outputs, tram seat-kilometers (η_5) and bus seat-kilometers (η_6), have a negative sign, implying that the network complexity has a lower effect in higher levels of output. Similarly, the positive coefficient of the interaction of the latent variable with the network length (η_7) suggests that the network complexity has a relatively greater effect in larger networks. The positive sign of the squared term of the latent variable (η_2) can also be interpreted as an increasing marginal effect of complexity. While all these interpretations appear to be consistent with the idea of linking the latent variable to network complexity, we should recognize that alternative interpretations could equally be justified. The results however point to the fact that the time-invariant heterogeneity is not separable from observed production factors.

The results listed in Table 2 also indicate considerable variation across companies with regard to time-invariant heterogeneity. The significant effect of interaction terms of the latent variable with outputs suggest that the technological characteristics such as the economies of scope or rates of returns to scale and density show a considerable variation across different companies. These variations are ignored in the model with separability assumption. In principle, such variations can be also modeled with a random coefficients model with several random effects. However, considering an identical latent variable allows a more tangible interpretation of such variations by associating them to unobserved characteristics such as network complexity. For instance, considering the latent variable as an inverse measure of the network complexity, we can interpret the

negative coefficients of the output interactions as an indication that more complex networks have higher rates of economies of scale.

The inefficiency scores u_i are summarized in Table 3. The estimated values vary from 0.01 to about 0.62. The values of the mean and median technical inefficiency are fairly low amounting to about 8 per cent.[24] A simple calculation based on the estimated coefficients of γ_i and γ_i^2, indicates that the effect of heterogeneity is rather substantial: Considering the estimated coefficients in Table 2 (especially η_7), one standard deviation of γ_i is approximately equivalent to about 0.14 or 0.28 depending on the model. These results suggest that the effect of time-invariant heterogeneity on inputs (and costs) is significantly greater than the average estimated inefficiencies. Moreover, in the model with separability assumption the coefficients of γ_i and γ_i^2, are significantly smaller suggesting that the effect of unobserved heterogeneities could be underestimated.

Table 3

5.0 Conclusions

Modeling unobserved heterogeneity in stochastic frontier literature is often based on certain assumptions about separability from observed production factors. Such separability assumptions can be restrictive in the context of transport networks, in which the entire production process is organized within given network structures entailing unobserved characteristics such as complexity and shape. This paper proposes a random

[24] For comparison purposes, we also estimated a "classical" model for panel data proposed by Pitt and Lee (1981) that considers any unobserved firm-specific heterogeneity as inefficiency. As expected, the values of technical inefficiency are higher and have more dispersion than those emerging from our models.

coefficient stochastic frontier model that allows for non-separability between unobserved time-invariant factors and observables.

An input distance function is used to examine the technical efficiency of a sample of Swiss and German urban transit companies. The results suggest that the estimated distance function could be a reasonable fit to the observed data. The estimated input and output elasticities have the correct sign and magnitude. The statistical tests favor the presence of considerable network heterogeneity and reject the separability assumption. The estimated scale elasticities indicate that the median company operates under both economies of density and scale. The analysis indicates that while the first-order coefficients of the distance function are not sensitive to the separability assumption, the second-order terms could differ significantly across the models. This is especially important in estimating the variation of technological properties such as returns to scale with output and network characteristics. In these cases, the proposed model can be used to relax the separability assumption, while allowing a possible association between unobservable factors and tangible structural characteristics such as network complexity.

In general, the results indicate considerable variation across companies in the marginal impact of the observed input and outputs, suggesting that the unobserved characteristics of the network structure play a crucial role in transport services. Thus, the proposed model can improve the estimates taking into account different unobserved network complexities. Finally, the results suggest that the effect of time-invariant heterogeneity could be greater than the estimated inefficiencies. This result underlines the possibility of substantial errors in the measurement of productive efficiency. Along with previous empirical studies, the present analysis confirms that the direct use of

benchmarking results in regulation could have significant and possibly undesired financial consequences for the regulated companies. Therefore, the benchmarking results should not be directly applied to define the tariffs applied to individual companies. However, the results can be used as an instrument to minimize the information asymmetry between the regulator and companies.

6.0 References

Aigner, D., Lovell, C. A. K., and P. Schmidt (1977): 'Formulation and Estimation of Stochastic Frontier Production Function Models', *Journal of Econometrics*, 6, 21-37.

Alvarez, A., Arias, C., and W. Greene (2004): 'Accounting for Unobservable in Production Models: Management and Inefficiency', Working Paper E2004/72, Fundacion Centro Estudios Andaluces.

Atkinson, S. E., and D. Primont (2002): 'Stochastic Estimation of Firm Technology, Inefficiency and Productivity Growth Using Shadow Cost and Distance Functions', *Journal of Econometrics*, 108(2), 203-225.

Bagdadioglu, N., and T. Weyman-Jones (2008): 'Panel Data Stochastic Frontier Analysis for Energy Network Regulation' Working Paper presented at Empirical Methods in Energy Economics, CEPE ETH Züruch, August 2008.

Battese, E., and T. J. Coelli (1988): 'Prediction of Firm Level Technical Inefficiencies with a Generalised Frontier Production Function and Panel Data', *Journal of Econometrics*, 38, 387-399.

Battese, G. E., and T. J. Coelli (1992): 'Frontier Production Function, Technical Efficiency and Panel Data: with Application to Paddy Farmers in India', *Journal of Productivity Analysis*, 3, 153-169.

Boitani, A., and C. Cambini (2002): 'Il Trasporto Pubblico Locale. Dopo la Riforma i Difficili Albori di un Mercato', *Mercato Concorrenza Regole*, 1, 45-72.

Cambini, C., and M. Filippini (2003): 'Competitive Tendering and Optimal Size in the Regional Bus Transportation Industry', *Annals of Public and Cooperative Economics*, 74(1), 163-182.

Coelli, T. J., and S. Perelman (1999): 'A comparison of Parametric and Non-Parametric Distance Functions: with Application to European Railways', *European Journal of Operations Research*, 117, 326-39.

Coelli, T. J., and S. Perelman (2000): 'Technical Efficiency of European Railways: A Distance Function Approach', *Applied Economics*, 32, 1967-1976.

Coelli, T. J. (2002): 'On the Econometric Estimation of the Distance Function Representation of a Production Technology', CEPA Working Paper, Centre for Efficiency and Productivity Analysis School of Economics, The University of Queensland, Queensland, Australia.

Colburn, C. B., and W. K. Talley (1992): 'A Firm Specific Analysis of Economies of Size in the U.S. Urban Multiservice Transit Industry', *Transportation Research Part B*, 3, 195-206.

Cornwell, C., Schmidt, P., and R. C. Sickles (1990): 'Production Frontiers with Cross-Sectional and Time-Series Variation in Efficiency Levels', *Journal of Econometrics*, 46(1-2), 185-200.

Dalen, D. M., and A. Gomez-Lobo (2003): 'Yardsticks on the Road: Regulatory Contracts and Cost Efficiency in the Norwegian Bus Industry', *Transportation*, 30, 371–386.

De Borger, B., Kerstens, K., and A. Costa (2002): 'Public Transit Performance: What Does one Learn from Frontier Studies?', *Transport Reviews*, 22(1), 1-38.

Demsetz, H. (1968): 'Why Regulate Utilities?', *Journal of Law and Economics*, 11, 55–65.

Estache, A., Rossi, M. A., and C. A Ruzzier (2004): 'The Case for International Coordination of Electricity Regulation: Evidence from the Measurement of Efficiency in South America', *Journal of Regulatory Economics*, 25(3), 271-295.

Färe, R., and D. Primont (eds.) (1995): *Multi-Output Production and Duality: Theory and Applications*, Kluwer Academic Publishers, Boston.

Farsi, M., Filippini, M., and W. Greene (2005): 'Efficiency Measurement in Network Industries: Application to the Swiss Railway Companies', *Journal of Regulatory Economics*, 28(1), 69–90.

Farsi, M., Fetz, A., and M. Filippini (2006a): 'Economies of Scale and Scope in Local Public Transportation', CEPE Working Paper No. 48, Centre for Energy Policy and Economics (CEPE), Zurich.

Farsi, M., Filippini, M., and M. Kuenzle (2006b): 'Cost Efficiency in Regional Bus Companies', *Journal of Transport Economics and Policy*, 40(1), 95–118.

Filippini, M., and R. Maggi (1992): The Cost Structure of the Swiss Private Railways, *International Journal of Transport Economics*, 19, 307-327.

Greene, W. (2004): 'Interpreting Estimated Parameters and Measuring Individual Heterogeneity in Random Coefficient Models', Department of Economics, Stern School of Business, Working Paper No. 04-08, New York University.

Greene, W. (2005a): 'Reconsidering Heterogeneity in Panel Data Estimators of the Stochastic Frontier Model', *Journal of Econometrics*, 126(2), 269–303.

Greene, W. (2005b): 'Fixed and Random Effects in Stochastic Frontier Models', *Journal of Productivity Analysis*, 23, 7-32.

Greene, W. (2007): *LIMDEP* 9.0, Econometric Modeling Guide, Volume 2, Econometric Software, Inc., 2007, Plainview, NY, USA

Hensher, D., and J. Stanley (2003): 'Performance-Based Quality Contracts in Bus Service Provision', *Transportation Research Part A*, 37(6), 519-538.

Hensher, D. (ed.) (2007): *Bus Transport: Economics, Policy and Planning*, Research in Transportation Economics, Volume 18, Elsevier, Oxford, UK.

Heshmati, A., and S. C. Kumbhakar (1994): 'Farm Heterogeneity and Technical Efficiency: Some Results from Swedish Dairy Farms', *Journal of Productivity Analysis*, 5(1), 45-61.

Jamasb, T., and M. Pollitt (2003): 'International Benchmarking and Regulation: an Application to European Electricity Distribution Utilities', *Energy Policy*, 31, 1609-1622.

Klemperer, P. (1999): 'Auction Theory: a Guide to the Literature', *Journal of Economic Surveys*, 13(3), 227-86.

Kumbhakar, S. C. (1991): 'Estimation of Technical Inefficiency in Panel Data Models with Firm- and Time-Specific Effects', *Economics Letters*, 36, 43-48.

Laffont, J. J., and J. Tirole (eds.) (1993): *A Theory of Incentives in Procurement and Regulation* The MIT Press, Cambridge, MA.

Lovell, C. A. K., Richardson, S., Travers, P., and L. L Wood (1994): 'Resources and Functioning: a New View of Inequality in Australia', in Eichhorn, W. (ed.) *Models and Measurement of Welfare and Inequality*, Springer Verlag, Berlin, 787-807.

O'Donnell C. J., and T. J. Coelli (2005): 'A Bayesian Approach to Imposing Curvature on Distance Functions', *Journal of Econometrics*, 126, 493–523.

Pitt, M., and L. Lee (1981): 'The Measurement of Sources of Technical Inefficiency in Indonesian Weaving Industry', *Journal of Development Economics*, 9, 43-64.

Saal, D. S., Parker, D., and T. Weyman-Jones (2007): 'Determining the Contribution of Technical Change, Efficiency Change and Scale Change to Productivity Growth in the Privatized English and Welsh Water and Sewerage Industry: 1985-2000', *Journal of Productivity Analysis*, 28, 127-139.

Schmidt , P., and R. Sickles (1984): 'Production Frontiers and Panel Data', *Journal of Business and Economics Statistics*, 2, 367–374.

Shleifer, A. (1985): 'A Theory of Yardstick Competition', *Rand Journal of Economics*, 16(3), 319-327.

Sickles, R.C., Good, D. H., and L. Getachew (2002): 'Specification of Distance Functions using Semi- and Non-parametric Methods with an Application to the Dynamic Performance of Eastern and Western European Air Carriers', *Journal of Productivity Analysis*, 17(1-2), 133–155.

Toner, J. P. (2001): 'The London Bus Tendering Regime. Principles and Practice', Presented at the VII International Conference on Competition and Ownership in Land Passenger Transport, Molde, Norway.

Tsionas, E. G. (2002): 'Stochastic Frontier Models with Random Coefficients', *Journal of Applied Econometrics*, 17, 127-147.

Viton, P. A. (1992): 'Consolidations of Scale and Scope in Urban Transit', *Regional Science and Urban Economics*, 22(1), 25-49.

Viton, P. A. (1993): 'How Big Should Transit Be? Evidence on the Benefits of Reorganization from the San Francisco Area', *Transportation*, 20, 35-57.

Appendices

Table 1: Summary statistics for Germany and Switzerland

Variable	**Obs**	**Mean**	**Min**	**Max**	**Obs**	**Mean**	**Min**	**Max**
German = GE; Swiss=CH	GE	GE	GE	GE	CH	CH	CH	CH
Covered population	616	366,709	40,800	164,2000	91	285,215	76,381	421,802
Number of employees	616	978	30	3996	91	953	76	2798
Network length tram in km	616	49	3	155	91	32	8	110
Network length bus in km	616	465	5	2653	91	139	42	362
Number trams	616	118	2	755	91	128	12	432
Number buses	616	135	2	470	91	167	30	314
Tram-km in 1000 km	616	5664	61	34,363	91	6,111	398	20,518
Bus-km in 1000 km	616	7211	86	28,519	91	8,121	1,525	18,438
Seat-km tram in 1000 km	616	96,4943	5000	6,187,000	91	847,835	37387	2,926,006
Seat-km bus in 1000 km	616	584,293	4000	2,303,000	91	974,580	121,443	2,283,553
Area in km^2	616	171	21	405	91	169	90	275

Table 2: Distance function estimation results

Variable	Parameter	Random parameter model with separable unobserved heterogeneity		Random parameter model with non-separable unobserved heterogeneity	
		Coefficient	**Standard error**	**Coefficient**	**Standard error**
Constant	α_i	-0.090*	0.008	0.031*	0.008
Ln(x_2/x_1)	α_{CT}	0.191*	0.007	0.243*	0.007
Ln(x_3/x_1)	α_{CB}	0.365*	0.012	0.357*	0.013
Ln(x_2/x_1)^2	α_{CTCT}	-0.051*	0.016	-0.060*	0.015
Ln(x_3/x_1)^2	α_{CBCB}	0.067*	0.028	0.124*	0.023
Ln(x_2/x_1)*ln(x_3/x_1)	α_{CBCT}	0.139*	0.014	0.098*	0.012
lny_1	β_T	-0.334*	0.006	-0.333*	0.006
lny_2	β_B	-0.485*	0.007	-0.456*	0.007
lny_1^2	β_{TT}	-0.113*	0.011	-0.110*	0.012
lny_2^2	β_{BB}	-0.174*	0.018	-0.179*	0.020
lny_1*lny_2	β_{BT}	0.114*	0.014	0.091*	0.015
Ln(x_2/x_1)*lny_1	δ_{TCT}	0.092*	0.013	0.086*	0.013
Ln(x_2/x_1)*lny_2	δ_{TCB}	-0.044*	0.014	-0.017	0.015
Ln(x_3/x_1)*lny_1	δ_{BCT}	-0.004	0.018	0.054*	0.017
Ln(x_3/x_1)*lny_2	δ_{BCB}	0.007	0.018	-0.084*	0.019
Trend	α_t	0.022*	0.001	0.022*	0.001
lnz_1	α_Z	-0.049*	0.006	-0.032*	0.006
lnz_1^2	α_{ZZ}	0.010	0.013	-0.033*	0.014
lnz_1*ln(x_2/x_1)	α_{ZT}	0.159*	0.010	0.138*	0.009
lnz_1*ln(x_3/x_1)	α_{ZB}	-0.119*	0.014	-0.109*	0.015
lnz_1*lny_1	α_{ZCT}	-0.122*	0.009	-0.131*	0.009
lnz_1*lny_2	α_{ZCB}	0.188*	0.009	0.206*	0.010
$\sigma = \sqrt{\sigma_u^2 + \sigma_u^2}$		0.123*	0.004	0.121*	0.004
$\lambda = \sigma_u / \sigma_v$		1.927*	0.225	2.322*	0.284
	Coefficients related to latent heterogeneity				
γ_i	η_1	0.136*	0.004	0.277*	0.008

γ_i ***ln(x_2/x_1)**	η_3			0.125*	0.010
γ_i ***ln(x_3/x_1)**	η_4			-0.130*	0.015
γ_i ***lny_1**	η_5			-0.021*	0.010
γ_i ***lny_2**	η_6			-0.023*	0.010
γ_i ***lnz_1**	η_7			0.024*	0.009
$\gamma_i * \gamma_i$	η_2	0.055*	0.006	0.093*	0.011
Wald Test H_0 : $\eta_3 = \eta_4 = \eta_5 = \eta_6 = \eta_7 = 0$ Chi-squared = 526.95 p-value = 0.000 **H_0 is rejected**					

Notes: The coefficient reported for each random parameter is the mean; (a) we report estimates of SD of normal distribution of random parameters. (*) indicates significance at the 5 per cent level.

Table 3: Descriptive statistics of inefficiency estimates

	Model 1 with separability assumption	**Model 2 with non-separability assumption**
Number of Observation	707	707
Mean	0.084	0.085
Std. Dev	0.053	0.057
Min	0.012	0.012
Median	0.071	0.069
Max	0.617	0.601