NHS reforms and efficiency of hospital services: a stochastic distance function approach.[1]

ALESSANDRA FERRARI

Department of Economics, University of Reading.

1. Introduction

The last two decades have witnessed major changes in the organisation of health services. Various factors, among which the aging of the population and technical progress, led to growing demand levels and expectations; then to discontent, cost increases and general financial crisis. In general terms this scenario has characterised (and still characterises) many European countries: thus the need to reform the system. The UK was one of the first European countries to introduce a reform, with the White Paper *Working for Patients* (December 1989) and the NHS and Community Care Act (June 1990, effective from April 1991) that changed many key features of the health system. As regards hospital services a clear distinction was introduced between purchasers (District Health Authorities and GP fund-holders) and providers (hospital trusts). By the acquisition of trust status hospitals were given a lot more autonomy in the management of their resources (Bartlett and Le Grand, 1994); on the other hand, however, their budget-based, cost reimbursement funding system was to be replaced by a contractual system: hospitals would have to sell their services to the purchasers via contracts on the so-called internal market. The idea was that competition for contracts would give hospitals incentives to efficient behaviour.

[1] Useful comments were received from Wiji Arulampalam, Simon Burke, John Cubbin, Dennis Leech and Tom Weyman-Jones, and from David Spiegelhalter as discussant of the paper at the RSS 2003 one-day meeting on "Performance Monitoring and Surveillance". Usual disclaimers apply.

Even though the "quasi-market" was formally eliminated in 1997, its main features remained in place, other countries are now following in the same direction, and the debate about the efficiency and effectiveness of this kind of reform is a topical policy issue. Not surprisingly it has generated a lot of scientific interest, but this has been mainly of a theoretical nature, on the efficiency properties of different contracts or the existence of competitive conditions on the market; not as much has been done instead from an empirical point of view, not least because of difficulties in getting the relevant data.

Soderlund *et al.* (1997) estimated a classical linear regression model on a sample of NHS hospitals in England for the years 1992 to1994; this revealed a general productivity improvement whose association with the changes to trust status remained however unsure.

Due to its easier availability, others have used the acute Scottish hospitals data set used in this paper[2]. For example Scott and Parkin (1995) used it for 1992/93 to estimate a translog cost function which highlighted the prevalence of constant returns to scale and economies of scope between different kinds of outputs (mainly inpatients and outpatients). Some specifical analysis of efficiency was performed by Maniadakis *et al.* (1997, 1999), who used Data Envelopment Analysis (DEA) to calculate Malmquist indexes of total factor productivity (TFP) for the period 1991/92-1995/96. They conclude a general improvement in TFP, mainly attributable to shifts of the frontier, but a worsening of the quality level[3]. However, due to its deterministic, non-parametric nature DEA suffers from a few relevant shortcomings, which are well

[2] ISD Scotland, Scottish Health Service Costs, NHS in Scotland.

[3] Measured as the survival rate 30 days after discharge.

documented in the literature[4]. The aim of this paper is to analyse the changes in technical efficiency and performance of hospitals during the years of the reform by means of a stochastic distance function. This contributes to the literature under the following respects. As opposed to DEA, the stochastic parametric approach allows the statistical testing of hypotheses, quantifying the reliability of the results. It also allows the analysis of the characteristics of the production process that a non-parametric method by definition does not identify. Furthermore the chosen model is a stochastic distance function (Coelli and Perelman, 1996) for technical efficiency, which is a frontier model as opposed to the classical linear regression one; as will be seen in Section 2, the frontier model specifically separates the noise in the data from the estimation of inefficiency, the latter being the aim of the exercise. The choice of a distance function form is due to the multiple output nature of the production process that rules out the direct estimation of a production function.

Data for the whole of the NHS unfortunately are not available. The analysis is therefore restricted to Scotland, on a sample of 52 acute hospitals observed between 1991/92 and 1996/97, thus covering the whole duration of the reform.

The paper is structured as follows. The estimation of stochastic frontiers and the distance function model are discussed in Section 2. Section 3 describes the data set. A first pooled model is estimated in Section 4, and a model with changing parameters in Section 5. General conclusions are in Section 6. The parameters' estimates are reported in Appendices 1 and 2; finally Appendix 3 reports the analysis of the residuals.

[4] General discussions on DEA can be found in Fried et al (1993); Charnes, Cooper, Lewin and Seiford (1994); Coelli, Rao and Battese (1998); Cooper, Seiford and Tone (2000). A specific analysis of its application to the hospital sector is in Parkin and Hollingsworth (1997).

2. The stochastic frontier and the distance function

Following Debreu (1951) and Farrell (1957) seminal papers, the efficiency of a firm can be defined and measured in terms of the distance of its actual performance from a frontier. If this frontier is the production function, i.e. the maximum attainable output from a given set of inputs, the distance will measure technical inefficiency. More formally, and in an output perspective, if there are N firms that use a vector $x \in R_+^K$ of inputs to produce a vector $y \in R_+^M$ of outputs then

$$P(x) = \{y : (x, y) \ is \ feasible\}$$

is the *output set*, i.e. all the levels of output that can be produced using a given level of inputs, whether efficient or not. The isoquant and the efficient subset are

$$IsoqP(x) = \{y \in P(x), \vartheta y \notin P(x), \vartheta(1, \infty)\}$$
$$EffP(x) = \{y \in P(x), y' \notin P(x), y_m' \geq y_m \forall m, y'_r > y_r \ \ for \ some \ r\}$$

and they identify respectively the boundary of the output set, defined in terms of radial expansions of the output points within it, and the frontier. As shown in Figures 1 and 2, in the case of a well behaved, continuously differentiable production technology the isoquant and the efficient subset coincide, whereas they do not with a piecewise linear frontier as the one estimated with DEA.

Following Shephard (1953, 1970) the distance function (i.e. the measure of (in)efficiency) is the radial expansion measure for the output vector(s) in order to reach the frontier. The distance function is defined as

$$D_o = \min\left\{\vartheta : \frac{y}{\vartheta} \in P(x)\right\} \qquad (1)$$

where $0 < D_o \leq 1$. If $D_o = 1$ the observation lies on the frontier, if $D_o < 1$ it lies below it and a radial expansion of $1/\vartheta$ of the outputs is necessary to reach it. The distance

function is homogeneous of degree +1 and weakly monotonically increasing in outputs, and it is invariant to changes in the units of measurement.

Fig.1: Piece-wise linear frontier, output maximisation case with one input and two outputs (y1 and y2).The isoquant is line ABCD, the efficient subset (the frontier) is BC, the (in)efficiency of point b is the distance from the frontier segment bb'.

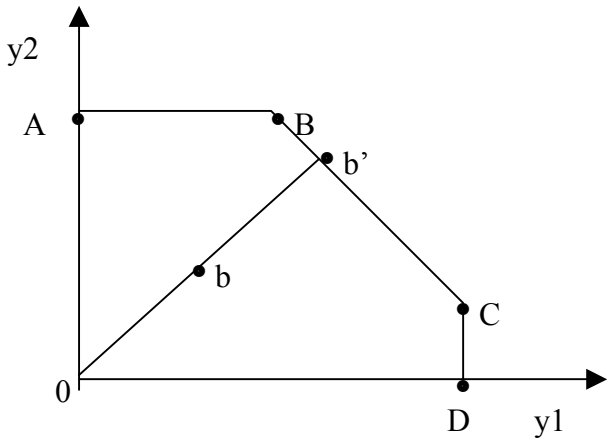
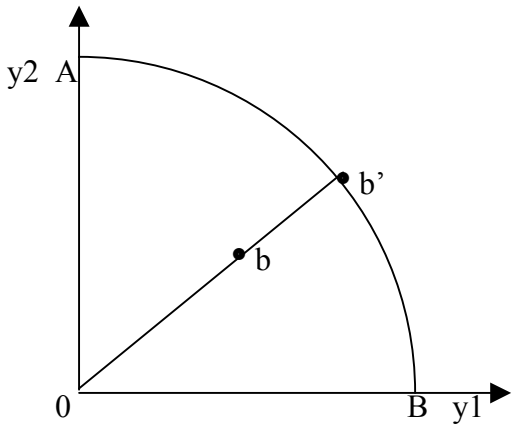


Fig2: Continuously differentiable frontier, output maximisation case with one input and two outputs (y1 and y2). The isoquant and the efficient subset (the frontier) coincide on line AB. The (in)efficiency of point b is the distance from the frontie, segment bb'.



Coming to estimation, the econometric stochastic frontier model was introduced at the same time by Aigner *et al* (1977), Battese and Corra (1977) and Meeusen and Van

den Broek (1977). In the single equation - cross sectional case, a production frontier is usually estimated as

$\ln y_i = \alpha + \beta' \ln x_i + \varepsilon_i$ (2)

where

$\varepsilon_i = v_i - u_i$

is a composite error term in which

$v_i \sim N(0,\sigma_v^2)$ is the stochastic component and

$u_i = -\ln D_o$ is the efficiency measure.

The efficiency measure u_i must come from a positively skewed, non negative distribution; for instance if this is a half normal (which will be used later), then

$u_i = |U_i|$

$U_i \sim N(0,\sigma_u^2)$

Due to the presence of a composite error term OLS gives consistent but inefficient parameters' estimates, and the use of ML is to be preferred if the distribution of u_i is known or an assumption can be made about it[5].

The influence of the inefficiency component can be measured by a parameter $\gamma = \sigma_u^2/\sigma^2$, where $\sigma^2 = \sigma_u^2 + \sigma_v^2$ (Battese and Corra, 1997). The significance of γ can be tested with an LR test which, if the null hypothesis H_0: $\gamma = 0$ is true, follows a mixed χ^2 distribution. If the null hypothesis is true and inefficiency is not significant, the model is equivalent to a standard "average" production function, and its log-likelihood is the same as that of the linear model estimated by OLS.

[5] Comprehensive reviews on the estimation of stochastic frontiers can be found in Greene (1997) and Kumbhakar and Lovell (2000).

As (2) can be estimated only for the single output case, an alternative model has been proposed by Coelli and Perelman (1996) to deal with the multiple outputs case. The idea is to directly express (1) as a function of the *K* inputs and *M* outputs of each of the *N* firms. Using a log-linear translog function specification[6] this is:

$$\ln D_{oi} = \alpha_0 + \sum_{m=1}^{M} \alpha_m \ln y_{mi} + \frac{1}{2} \sum_{m=1}^{M} \sum_{n=1}^{M} \alpha_{mn} \ln y_{mi} \ln y_{ni} + \sum_{k=1}^{K} \beta_k \ln x_{ki} +$$
$$+ \frac{1}{2} \sum_{k=1}^{K} \sum_{l=1}^{K} \beta_{kl} \ln x_{ki} \ln x_{li} + \sum_{k=1}^{K} \sum_{m=1}^{M} \delta_{km} \ln x_{ki} \ln y_{mi} \qquad (3)$$

$$i = 1,....., N$$

Linear homogeneity in outputs of D_o implies that

$$D_o(x, \omega y) = \omega D_o(x, y) \quad \forall \omega > 0$$

If we choose any of the *M* outputs, say the *M*th one, and set $\omega = 1/y_M$ so that

$$D_o(x, y/y_M) = D_o(x, y)/y_M \qquad (4)$$

we can impose linear homogeneity on (3) that becomes

$$\ln\left(\frac{D_{oi}}{y_{Mi}}\right) = \alpha_0 + \sum_{m=1}^{M-1} \alpha_m \ln y^*_{mi} + \frac{1}{2} \sum_{m=1}^{M-1} \sum_{n=1}^{M-1} \alpha_{mn} \ln y^*_{mi} \ln y^*_{ni} + \sum_{k=1}^{K} \beta_k \ln x_{ki} +$$
$$+ \frac{1}{2} \sum_{k=1}^{K} \sum_{l=1}^{K} \beta_{kl} \ln x_{ki} \ln x_{li} + \sum_{k=1}^{K} \sum_{m=1}^{M-1} \delta_{km} \ln x_{ki} \ln y^*_{mi} \qquad (5)$$

$$i = 1,...., N$$

where $y^*_m = y_m / y_M$. When $m = M$, $y_m = y_M$ the ratio is equal to one and so its log is equal to zero and all the terms involving the M-th output disappear from the equation. For simplicity, let TL(.) represent the right hand side of the translog function in (5). This can be estimated by noting that

$$\ln(D_{oi}/y_{Mi}) = TL(.)$$

is equivalent to

[6] A more parsimonious specification like the Cobb-Douglas cannot be used because, apart from other restrictions, it is not concave in the output dimensions.

$$-\ln y_{Mi} = TL(.) - \ln D_{oi} \quad (6)$$

Adding the stochastic component[7] $v_i \sim N(0,\sigma^2_v)$ and setting $\ln D_{oi} = -u_i$, equation (6) becomes

$$\ln y_{Mi} = -TL(.) + v_i - u_i \quad (7)$$

This can be now estimated as a usual production frontier, by regressing (the log of) one output on (the logs of) the inputs and (the logs of) the outputs ratios. Note that the coefficients of a production frontier correspond to the negative of the coefficients of a distance function.

3. The data

The data are a sample of 52 acute hospitals in Scotland in the years 1991/92 to 1996/97 (from now on referred to as 1992 and 1997 respectively), that make a panel data set of 312 observations. These data were obtained from the Scottish Health Service Costs statistics.

As regards the definition of the input and output variables, the following choices have been made[8]. Output is usually measured as the total number of cases treated, an intermediate measure given the difficulty of measuring the final improvement in health. Cases are however very heterogeneous, and for this reason they are usually divided into various specialty (or casemix) categories, which qualify the hospital as a multiple-output unit. A trade-off therefore exists between the aim of preserving this heterogeneity and the degrees of freedom of any estimation. The use of index

[7] This is the usual Gaussian error of a regression, which makes the model a stochastic, instead of deterministic, frontier.
[8] See in particular McGuire *et al*, 1983; Tatchell, 1983; McGuire, 1985; Butler, 1995.

numbers can overcome the problem as long as one can define weights that correctly represent the differences between cases.

For this paper, the following choice was made. The many output categories have been summarised in two indexes: one for the inpatients and one for the outpatients, day patients and day cases. The main difference between the two categories is that in the former patients spend several days in the hospital and in the latter no more than one day, sometimes without even using a bed (and the staff associated with it). As substitution between the two kinds of services could have taken place, it was preferred to keep them separate rather than summarising everything in one output measure.

For the weights, a measure of the average costliness of a case in each category of treatment has been calculated. The assumption is that more difficult illnesses are more input demanding than the less serious ones, and will therefore have a higher average cost[9]. Two adjustments have been made to the simple average cost per case. First, in order not to bias the weights with some measure of inefficiency of each hospital, all averages are calculated for the whole of Scotland. Second, as average costs change over time if inefficiency changes, they have been normalised each year to sum to 1. In detail, define

q_{jit} as the total number of cases treated in category j by hospital i at time t, and

c_{jit} as the average cost per case in category j at hospital i at time t; then

$c_{.jt} = \frac{1}{N}\sum_{i=1}^{N} c_{jit}$ is the average cost (across hospitals) per case in category j at time t

The weight for each category is calculated as

$$w_{jt} = \frac{c_{.jt}}{\sum_{j=1}^{J} c_{.jt}}$$

[9] Some might require the use of particular equipment, and/or more staff time, as well as a longer time spent in the hospital, which in turn implies more inputs use and therefore a higher cost.

and so finally the index for each hospital in each year is

$$y_{it} = \sum_{j=1}^{J} w_{jt} q_{jit}$$

As the two main output categories were kept separated, two final output indexes were calculated as above, which are:

y_1= index of inpatients

y_2= index of outpatients, day patients and day cases.

Finally, 5 variables identify the inputs:

x_1= total capital charges (£000)

x_2= medical staff FTE (full time equivalent);

x_3= nursing staff FTE;

x_4= other staff FTE;

x_5= total number of beds.

Capital[10] is measured in £000, and it is deflated using the "Hospital and Community Health Services pay and price inflation values". The "other staff" input includes professional, technical, administrative, clerical and all other staff. The descriptive statistics of the data are reported in Table 1.

[10] This was the only available capital measure on the data set, and it comprises: a) depreciation on fixed assets; b) interest paid on money borrowed to finance any of the projects in a); c) a 6% return on capital (trusts only).

Table 1: Descriptive statistics of the inputs and outputs variables (standard deviation into brackets)

	Average 92-97		Average annual rate of growth
Inpatients index	141746	(125124)	0.3
Outpatients et al. index	30342	(28349)	9.0
Capital (£000)	1513	(1396)	1.6
Medical staff (FTE)	88	(87)	4.1
Nursing staff (FTE)	457	(360)	0.7
Other staff (FTE)	302	(256)	4.0
Beds	357	(272)	-2.7

4. The model

The first model estimated is a translog output distance function where (the log of) the index of outpatients, day patients and day cases is regressed over (the log of) the five inputs (x) and the outputs ratio (y* = y1/y2), plus five dummy variables to allow for a different intercept each year, and a dummy variable for teaching hospitals. A dummy variable for trust status could not be introduced because of the implicitly assumed

correlation with the inefficiency component (the issue is discussed more in detail in Section 5).

The model full specification therefore reads as

$$\ln y_{2it} = \alpha_0 + \alpha_1 \ln y^* + \alpha_{11} (\ln y^*)^2 + \sum_{k=1}^{5} \beta_k \ln x_{kit} + \\ + \sum_{k=1}^{5} \sum_{l=k}^{5} \beta_{kl} \ln x_{kit} \ln x_{lit} + \sum_{k=1}^{5} \delta_k \ln x_{kit} \ln y^* + \sum_{t=1}^{5} \zeta_t D_t + \xi D_{teach} + v_{it} - u_{it} \quad (8)$$

where

$i = 1,\ldots,N \quad and\ N=52$

$t = 1,\ldots,T \quad and\ T=6$

$v_{it} \sim NIID(0, \sigma^2_v)$

is the stochastic component, coming from a normal distribution. The inefficiency component u_{it} is modelled by allowing it to vary stochastically between hospitals but deterministically across time. Specifically, following Battese and Coelli (1992)

$u_{it} = u_i exp[-\eta(t-T)]$

$u_i = |U_i|$

$U_i \sim NIID(0, \sigma_u^2)$

where a value of $\eta>0$ (<0) implies increasing (decreasing) efficiency over time. A value of $\eta=0$ implies no time effect, and the hypothesis can be tested by means of an LR test. The choice of a half normal distribution for u_i was made after testing it against the more general specification of the truncated normal[11]. Finally

$\varepsilon_{it} = v_{it} - u_{it}$

is the composite error term.

[11] The details of the testing procedure and results are available from the author on request.

The choice of (8) was made after estimating and comparing four alternative models: a base model without any time effect (M1), one with a linear time trend, one with a quadratic time trend (M3) and the dummy variables model in (8) (M4). As all models are nested in one another, the comparisons were made by means of LR tests, and the results are reported in Tables 2 and 3. The tests sequence is the following: first test whether there is a time effect or not: M1 (the null hypothesis) is tested against M4 (dummy variables), M2 (quadratic time trend), and M3 (linear time trend). In all cases the null hypothesis has to be rejected (see Table 5.2), meaning that a time effect exists. Then the three models including a time variable are tested against one another. The null hypothesis of a linear model is rejected against the unrestricted quadratic time tend, which in turn is rejected as a null hypothesis against the dummy variables specification. More detail about the testing procedure is given in Appendix 4.

The estimations were carried out via maximum likelihood, using the software FRONTIER 4.1.

Table 2: log likelihood of models M1 to M4.

	$\mathcal{L}$	**# parameters**
M1	158.60	29
M2	164.80	30
M3	167.86	31
M4	179.13	34

Table 3: comparison of models M1 to M4 (# restrictions into brackets).

	LR score		Implication
M1 vs M2	12.4	(1)	H_0 (M1) rejected
M1 vs M3	18.52	(2)	H_0 (M1) rejected
M1 vs M4	41.06	(5)	H_0 (M1) rejected
M2 vs M3	6.12	(1)	H_0 (M2) rejected
M3 vs M4	22.54	(3)	H_0 (M3) rejected

Table 3 shows that a time effect should be included and that the dummy variables specification is to be preferred. The full results of the estimation of M4 are reported in Appendix 1, which shows a number of coefficients are not significant, and the main results are summarised in Table 4: the first two columns report the estimates of the intercept and dummy variables and the two log-likelihood values; the last two columns report the estimates of the elasticities and of the parameters γ and η. All elasticities are calculated at the sample mean and their significance is tested by means of an LR test that in all cases leads to reject the null hypothesis (i.e. all are significant)[12]. The inputs partial elasticities are reported first, followed by the elasticity of y_2 with respect to the outputs ratio $y^* = y_1/y_2$, then by the total input elasticity (or elasticity of scale) and finally by the elasticity of substitution between y_2 and y_1, whose calculation is detailed as follows.

Assume for ease of explanation that the estimated function has one input and two outputs (whose ratio is again denoted by y^*), as

$$\ln y_2 = a_1 \ln x + a_{11}(\ln x)^2 + b_1 \ln y^* + b_{11}(\ln y^*)^2 + c \ln x \ln y^* \qquad (10)$$

[12] The presence of the squared and interaction terms makes the translog prone to multicollinearity. As a consequence it is usually advisable to test for joint parameters' significance rather than relying on their individual ones.

Total differentiation of (10) is

$$\partial \ln y_2 = A\partial \ln x + B\partial \ln y*$$

or equivalently

$$\partial \ln y_2 = A\partial \ln x + B\partial \ln y_1 - B \ln y_2 \qquad (11)$$

where

$$A = \frac{\partial \ln y_2}{\partial \ln x} \quad and \quad B = \frac{\partial \ln y_2}{\partial \ln y*} = e_{y*} \qquad (12)$$

The elasticities with respect to y_2 alone can now be calculated as

$$\frac{\partial \ln y_2}{\partial \ln x} = \frac{A}{(1+B)} \qquad (13)$$

and

$$\frac{\partial \ln y_2}{\partial \ln y_1} = \frac{B}{(1+B)} \qquad (14)$$

Equation (13) is the elasticity of y_2 with respect to input x, so it should be >0. Equation (14) is the elasticity of y_2 with respect to y_1, i.e. it is a measure of the substitutability between the two outputs and it should be < 0. From (14) it follows that $-1 < B < 0$: lower absolute values of B will imply very little substitutability between the two outputs, and higher absolute values of B a higher substitutability.

Coming to the results, Table 4 shows that inefficiency is significant and significantly decreases over time: the estimated value of η is 0.09; this corresponds to an average rate of change of the distance function $d(lnD_o)/dt$ of around 2.45% per year (as approximated by the difference in the logs). The parameter for teaching status is positive but not significant. With the exception of capital, all inputs elasticities are

positive. Medical and nursing staff are the most productive inputs, whereas capital and beds are the least, with the former showing negative returns, a result difficult to interpret. The total elasticity of scale is 2.82, so the production function shows increasing returns to scale (at the sample means). Given the particular functional specification used, this measures the effect that an increase in inputs has on the output, *given the outputs ratio*: if the outputs ratio remains the same then a 1% increase in all inputs leads to an increase of 2.82% in *both* outputs. Not surprisingly the hypothesis of constant returns to scale is rejected (LR score 20.5, with 7 restrictions), but the result seems to be due only to the very high value of the partial elasticity of the medical staff.

The variable e_{y*} is the elasticity of y_2 with respect to the outputs ratio y*. This is a measure of the curvature of the frontier, called B in (12), and as expected it is a negative -0.67). This translates into an output substitutability of –2.075, meaning that a 1% increase in y_1 (the inpatients) leads to a more than proportional decrease in y_2 (the outpatients, day patients and day cases): as one might expect inpatients appear a lot more expensive, in terms of resource use, than outpatients, day patients and day cases.

Table 4: results from the estimation of M4 (time dummy variables).

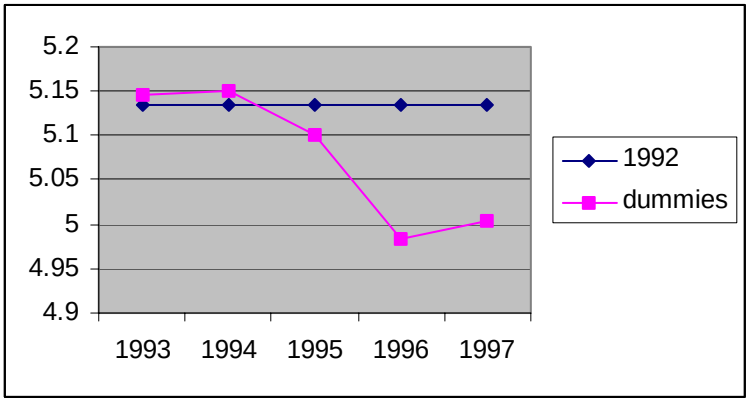
variable	estimate (t-statistic)	variable	Estimate (LR statistic)
α_0	5.13	e_{cap}	-0.01
	(6.52)*		(25.86)*
D_{93}	0.01	e_{med}	2.31
	(0.44)		(89.24)*
D_{94}	0.02	e_{nurs}	0.34
	(0.57)		(25.08)*
D_{95}	-0.03	e_{oth}	0.14
	(0.99)		(34.02)*
D_{96}	-0.15	e_{bed}	0.04
	(4.05)*		(23.82)*
D_{97}	-0.13	e_{y*}	-0.67
	(3.02)*		(201.13)*
D_{teach}	0.03	e_{tot}	2.82
	(0.69)		
		e_{y1}	-2.07
$\mathcal{L}$	179.13	γ	0.91*
			(308.44)
OLS $\mathcal{L}$	24.91	η	0.09*
			(26.86)

* = significant at 5% (or less);
** = significant at 10%.

Coming finally to the intercept dummies, their pattern is shown in Figure 3.

The figure shows that, starting from 1992, there is a mild increase in 1993 and 1994, whereas from 1995 the trend is markedly decreasing. If the dummies account for technological change then the above results would indicate a slowdown in productivity over time (at least for one of the outputs), especially after the first big change to trust status which takes place in 1994.

Figure 3: Time pattern estimated by the dummy variables model (8).



The overall picture would then be one of progressive worsening of productivity, which might in turn be the reason of the increase in technical efficiency (a lower frontier is easier to reach), and where one of the inputs consistently negatively affects production. Together with the negative capital elasticity, this scenario raises the doubt that the effect of time might not have been adequately captured. Another possibility is therefore explored: that not only the intercepts but all the parameters of the equation might have changed over time. This possibility is explored in the next section.

5. Testing for technological change

The results obtained from the estimation of M4 raised the suspicion that a pooled model might not be correct, and that all the parameters, and not just the intercepts might have changed over time. As the use of Chow tests for parameters stability is ruled out for lack of degrees of freedom (the translog has too many parameters to be estimated on a single cross section of 52 observations), an alternative approach is used instead. This consists of estimating several times the distance function with a time interaction dummy instead of the intercept dummies. In particular, a time dummy *d* is introduced, which takes a value of 1 for a particular year(s), and 0 else, and this is multiplied to all the variables in the translog distance function, as:

$$\begin{aligned}\ln y_{2it} &= \alpha_0 + \alpha_1 \ln y^* + \alpha_{11}(\ln y^*)^2 + \sum_{k=1}^{5} \beta_k \ln x_{kit} + \\ &+ \sum_{k=1}^{5}\sum_{l=1}^{4} \beta_{kl} \ln x_{kit} \ln x_{lit} + \sum_{k=1}^{5} \delta_k \ln x_{kit} \ln y^* + \lambda_0 d + \lambda_1 \ln y^* d + \lambda_{11}(\ln y^*)^2 d + \\ &+ \sum_{k=1}^{5} \rho_k \ln x_{kit} d + \sum_{k=1}^{5}\sum_{l=1}^{4} \rho_{kl} \ln x_{kit} \ln x_{lit} d + \sum_{k=1}^{5} \zeta_k \ln x_{kit} \ln y^* d + \xi D_{teach} + v_{it} - u_{it}\end{aligned} \qquad (15)$$

In (15) the x_{it}s, the y_{it}*s and D_{teach} are the same explanatory variables as in (8) and *d* is the time interaction dummy. When *d=0* the parameters of the function are the αs, βs and δs; when *d=1* they are the respective *(αs+λs), (βs+ρs) and (δs+ζs).*

The dummy is first set equal to 1 for 1992 (and 0 else), then for 1992 and 1993 (and 0 else) and so on. In this way 5 different distance functions are estimated, each with a different time effect which is captured by the parameters of the interaction dummy. The likelihood results of the five estimations of (15) are reported in Table 6. In each case the significance of the time interaction parameters is tested for by means of LR

tests against a restricted model with no time effects and as expected the null hypothesis is always rejected.

Table 6: Log-likelihood of the translog distance function with time interaction dummy.

	92	**92 - 93**	**92 - 94**	**92 - 95**	**92 - 96**
$\mathcal{L}$	177.99	218.54	190.93	191.68	170.19

Since the models are not nested in one another their comparison should be made on the basis of information criteria. In this case all the models have the same number of parameters, and so selection by minimisation of any standard information criterion is equivalent to selection of the model with the greatest maximised log likelihood. This happens when separating 1992 and 1993 from all other years, as the model has about 27 points of difference in the log-likelihood from its closest alternative. This therefore points to the fact that the parameters of the distance (and production) function might have changed after 1993. On the grounds of the Akaike[13] information criterion the model as specified in (15) is also to be preferred to the pooled one (M4).

Given the above, the model in which 1992 and 1993 are separated from the following years is analysed. The main results are shown in Table 7, whereas all the parameters' estimates are in Appendix 2. Appendix 3 reports the graphs of the true density function of the composite error term and that of the estimated residuals; their

[13] The Akaike information criterion is used to compare models that are non nested in one another. It is specified as $AIC = -2\mathcal{L} + 2n$ where $\mathcal{L}$ is the value of the maximised log-likelihood and n is the number of parameters. The preferred model is the one with the lowest AIC value.

comparison shows that the model fits the data well and rules out the presence of outliers.

Looking at Table 7, the γ parameter is significant (LR test is 337.62) meaning that so is inefficiency. Very interestingly, however, this time η is not: the result of the LR test (2.46) leads not to reject the null hypothesis that $\eta = 0$, so that no significant difference in (in)efficiency appears to have taken place over time.

The most notable difference between the two time periods is the change in e_{y1}, the elasticity of substitution between y_2 and y_1: the absolute value increases by a 60%, meaning that the opportunity cost of treating someone as an inpatient over time becomes a lot higher. A pattern therefore is revealed towards treating patients more and more on a day basis, as day patients/cases if not directly as outpatients. This is confirmed when looking at Table 1, that shows a very big rate of increase in the value of y_2 as opposed to a relatively small increase in that of y_1.

As regards the inputs elasticities, only two variables improve their performance after 1994 (namely the nursing staff and the beds), and the other three lower it. Whether this is due to a specific change in the technology involving each input or just to a change in its levels can be revealed by testing for the significance of the respective dummy variable parameters. The improvement in productivity of the beds input then appears to be a consequence of the reduction in the levels of the variable, well known also to the general public via the news. This is also consistent with the aforementioned trend towards day-based treatment, as beds would in that case be used more intensively.

This nursing staff variable also shows the lowest rate of increase in levels over time, although this appears not to be the only reason of its improved productivity.

Table 7: results of the estimation of the time-interaction dummy variable model allowing for technical change (equation (15)).

Variable	Estimate		
$\mathcal{L}$	218.54		
OLS $\mathcal{L}$	49.73		
γ	0.955*		
η	0.02		
	1992-1993	**LR score (when D=0)**	**1994-1997 (LR score)**
e_{cap}	0.028*	(17.56)	0.005*
e_{med}	1.049*	(46.38)	0.672*
e_{nurs}	0.156*	(17.6)	0.254*
e_{oth}	0.334*	(20.94)	0.121*
e_{bed}	0.043*	(41.04)	0.289*
e_{y*}	-0.493*	(232.3)	-0.613*
e_{tot}	1.611		1.341
e_{y1}	-0.973		-1.582

* = significant at 5% (or less).
** = significant 1 at 10%.

The reduced elasticities of capital and other staff are taken to be a direct consequence of the reform. As regards capital, the increase in the variable levels could be due to the investment in information technology that hospitals made in order to deal with the new contracting issues (Fattore, 1999). As this activity is not directly linked to the treatment of patients (the output variable) this might explain the reduced productivity

of the input. However, increased capital levels are also the consequence of accountancy changes related to the change to trust status, which made the hospitals owners of their assets, so concluding a definite problem of overcapitalisation would be misleading. As the data did not offer any other measure of capital but the one used, no further detail is available and the result has to be taken with caution.

Similarly, the "other staff" variable increases in level and its elasticity decreases from 0.33 to 0.12. One reasonable explanation is the increased administrative staff made necessary to deal with the new contracting issues. Another possibility is that the lower increase in nursing staff might have led to the transfer of some of their duties over cheaper but less qualified (and therefore less efficient) staff. A pattern towards the use of cheaper labour inputs in Scottish hospitals was revealed by others (Gray *et al.*, 1986), and this might have been reinforced by the financial concerns of the reform.

Finally, the difference in the intercepts indicates an improvement in average productivity, that is a shift upwards of the frontier. Considering that the shape of the frontier has changed, the higher intercept could indicate that the improvement is mainly in the production of the dependent variable, i.e. again y_2.

The fact that 1994 is the year of the first big change to trust status naturally leads to think of that as the reason behind the structural break. The relevance of trust status in explaining changes in technology and inefficiency is therefore analysed more in detail.

The correct approach to the analysis of the determinants of inefficiency is the estimation of a one-step model (Wang and Schmidt, 2002): this specifies the

inefficiency term as a function of some explanatory variables, and then simultaneously estimates its parameters and those of the frontier itself. One model of this kind is proposed by Battese and Coelli (1995) and can be estimated by FRONTIER 4.1 (Coelli 1996). However, no further details are given because the several, different attempts to estimate it all failed to converge to a maximum. More sophisticated models are currently discussed by the literature and will be worth exploring for future research. For this paper the interaction dummy specification remains the preferred one. Using the estimates of that model, the elasticities of trusts and non trusts are calculated and compared to one another, as shown in Table 8[14].

Table 8: partial elasticities of trusts and non trusts hospitals.

	Trusts	**Non trusts**
e_{cap}	-0.001	0.022
e_{med}	0.667	0.687
e_{nurs}	0.252	0.259
e_{oth}	0.121	0.128
e_{bed}	0.308	0.207
e_{y1}	-1.594	-1.506

Table 8 shows that the pattern revealed by Table 7 after 1994 seems to be more marked for the trusts sample than for the other hospitals, which confirms the hypothesis that the change in technology is related to the change in status. However, no significant link between trust status and efficiency can be detected: a t-test on the

[14] In particular these are calculated at the average sample values in 1994 and 1995, which are the years where a reasonable mix of trusts and non trusts exists. 1992 and 1997 in fact have 0 in one category, and 1993 and 1996 have 7 or less in one category.

equality of the mean inefficiency score, computed as $E[u_i / \hat{\varepsilon}_i]$[15] of trusts and non trusts is performed and the null hypothesis cannot be rejected (the p-value is 0.34).

Putting all the evidence together, the following general picture is disclosed. The main effect that the reform seems to have on hospitals is to change their technology. A structural break is detected in 1994, the year of the first trust wave, after which hospitals change not only the way in which they provide their services, but also the mix of services they provide. The opportunity cost of inpatients increases as hospitals tend to treat patients more and more on a day basis. This view is supported by the fact that both the number and the costliness of outpatients, day patients and day cases increase quite significantly, indicating a possible "swap" between the two categories of output considered. This could be the result of hospitals attempting to reduce their costs by reducing the length of stay, especially if the contracts constrained them to provide minimum levels of treatment (as it was the case especially with the widely used 'block contracts'). The involvement in the new contracting activity, and the financial concerns that this brings, also appear to translate into reduced inputs productivity. This at least seems to be true for the capital and other staff variables (whose levels increase over time), and for the medical staff. Nurses and beds instead improve their productivity, and are also associated with the lowest increases in levels (with the latter strongly negative in fact).

This increase in the day-basis treatment is also behind the shift upwards of the frontier. However, although technical inefficiency remains significant, it does not show any significant improvement. If a shorter length of stay raises concerns about

[15] See Greene (1997) for a discussion.

the quality and effectiveness of treatment, then the overall scenario might not be particularly optimistic.

6. Conclusion

The UK undertook in the 90's a major reform of their health system, introducing, among other things, an internal market for hospital services; the idea was that the competitive environment would improve the efficiency of hospitals' provision. The system was reformed again in 1997, but many of the new key features remained the same. Albeit being an interesting and relevant issue, the effectiveness of the internal market has not been extensively analysed by the empirical economic literature.

The aim of this paper was to analyse the changes in (technical) efficiency and performance of acute hospitals in Scotland during the years of the reform, from 1991/92 to 1996/97. The econometric tool was the estimation of a stochastic distance function. Different models were estimated, and the analysis led to finally choose one that allows all the technology parameters to change over time. This revealed a structural break associated with the change to trust status (which embodies the full working of the reform) after which hospitals changed both the way in which they provide their services and the kind of services they provide. In particular, the former showed as a lower productivity of most inputs, the latter as a trend towards the "quicker" treatment of patients on a day basis. No improvement in technical efficiency was detected instead. Consistently with other literature, this was interpreted as hospitals having to devote resources to new activities and concerns, which could however happen at the expenses of the effectiveness and quality of their services.

There are clearly limitations to this work, the first of which is the non-availability of a specific quality measure in the data set. Another is that the sample is relatively small, and a more robust analysis would ideally enjoy a higher number of observations, possibly covering the whole country. With these in mind, the general conclusion that can be drawn from this work is that the reform did not unambiguously produce all the expected beneficial effects.

References

Aigner, D., Lovell, C. A. K. and Schmidt, P. (1977). "Formulation and estimation of stochastic frontier production function models", *Journal of Econometrics*, 6, 21-37.

Bartlett, W. and Le Grand, J., (1994). "The performance of trusts", in Robison, R. and Le Grand, J. (1994), *Evaluating the NHS reforms*, The King's Fund Institute.

Battese, G. E. and Coelli, T. J. (1992). "Frontier production functions, technical efficiency and panel data: with application to paddy farmers in India", *Journal of Productivity Analysis*, 3, 153-169.

Battese, G. E. and Coelli, T. J. (1995). "A model for technical inefficiency effects in a stochastic frontier production function for panel data", *Empirical Economics*, 20, 325-332.

Battese, G. E. and Corra, G. (1977). "Estimation of a production frontier model: with application to the pastoral zone of eastern Australia", *Australian Journal of Agricultural Economics*, 21, 167-179.

Butler, J. (1995). *Hospital Cost Analysis*, Kluwer Academic Publishers.

Charnes, A., Cooper, W. W., Lewin and A., Seiford, L. M. (1994). *Data Envelopment Analysis: Theory, Methodology and Application*; Kluwer Academic Publishers.

Coelli, T. J. (1996). "A guide to FRONTIER version 4.1: a computer program for stochastic frontier production and cost function estimation", *CEPA working paper* 96/07, University of New England, Australia.

Coelli, T. J. and Perelman, S. (1996). "Efficiency measurement, multiple output technologies and distance functions: with application to European railways", *CREPP working paper* 96/05, Université de Liège.

Coelli, T. J., Rao, D. S. P. and Battese, G. E. (1998). *An Introduction to Efficiency and Productivity Analysis*; Kluwer Academic Publishers.

Cooper, W. W., Seiford, L. M. and Tone, K. (2000). *Data Envelopment Analysis. A Comprehensive Text with Models, Applications, References and DEA-Solver Software*; Kluwer Academic Publishers.

Debreu, G. (1951). "The coefficient of resource utilisation", *Econometrica* 19(3), 273-292.

DoH, (1989). *Working for Patients*, HMSO, London (Cm 555).

Farrell, M. J. (1957). "The measurement of productive efficiency", *Journal of the Royal Statistical Society*, Series A, General, 120 (3), 253-281.

Fattore, G. (1999). "Cost containment and health care reforms in the British NHS", in Mossialos, E. and Le Grand, J. (1999), *Health Care and Cost Containment in the European Union*, Ashgate, USA.

Fried, H. O., Lovell, C. A. K. and Schmidt, S. S. (1993). *The Measurement of Productive Efficiency- Techniques and Applications*, Oxford University Press.

Gray, A., McGuire, A. and Stuart, P. (1986). "Factor input in NHS hospitals", *Discussion Paper* n.2, University of Aberdeen.

Greene, W.H. (1997). "Frontier production functions", in Pesaran, M. H. and Schmidt, P. (1997), *Handbook of Applied Econometrics*, 2, Microeconomics, Blackwell.

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

Maniadakis, N. and Thanassoulis, E. (1997). "Changes in the productivity of a sample of Scottish hospitals: a cost index approach", *Research Paper* no. 288, Warwick Business School.

Maniadakis, N., Hollingsworth, B. and Thanassoulis, E. (1999). "The impact of the internal market on hospital efficiency, productivity and service quality", *Health Care Management Science*, 2, 75-85.

McGuire, A. and Westoby, R. (1983). "A production function analysis of acute hospitals", *Discussion Paper* n. 4, University of Aberdeen.

McGuire, A. (1985). "Methodological considerations of hospital production and cost functions: relationships to efficiency", *Discussion Paper* n.8; University of Aberdeen.

McGuire, A., Henderson, J. and Mooney, G. (1988). *The Economics of Health Care*, Routledge and Kegan, London;

Meeusen W. and Van den Broek, (1977). "Efficiency estimations from Cobb-Douglas production functions with composed error", *International Economic Review*, 18, 435-444.

Mossialos, E. and Le Grand, J. (1999). *Health Care and Cost Containment in the European Union*, Ashgate, USA.

National health Service in Scotland (1991-1997). *Scottish Health Service Costs*.

Parkin, D. and Hollingsworth, B. (1997). "Measuring production efficiency of acute hospitals in Scotland 1991-1994: validity issues in data envelopment analysis", *Applied Economics*, 29, 1425-1433.

Robison, R. and Le Grand, J. (1994). *Evaluating the NHS reforms*, The King's Fund Institute.

Scott, A. and Parkin, D. (1995). "Investigating hospital efficiency in the new NHS: the role of the translog cost function", *Health Economics*, 4, 467- 478.

Shephard, R. W. (1953). *Cost and Production Functions*, Princeton N.J.; Princeton University Press.

Shephard, R. W. (1970). *Theory of Cost and Production Functions*, Princeton N.J.; Princeton University Press.

Söderlund, N., Csaba, I., Gray, A., Milne, R. and Raftery, J. (1997). "Impact of the NHS reforms on English hospital productivity: an analysis of the first three years", *British Medical Journal*, 315,1126-1129.

Tatchell, M. (1983). "Measuring hospital output: a review of the case mix and service mix approaches", *Social Science and Medicine*, 17, 871-883.

Wang, H and Schmidt, P (2002). "One-step and two-step estimation of the effects of exogenous variables on technical efficiency levels", *Journal of Productivity Analysis*, 18, 129-144.

Appendix 1

Table A1 reports the results of the estimation of equation (8):

$$\ln y_{2it} = \alpha_0 + \alpha_1 \ln y^* + \alpha_{11} (\ln y^*)^2 + \sum_{k=1}^{5} \beta_k \ln x_{kit} +$$

$$+ \sum_{k=1}^{5} \sum_{l=1}^{4} \beta_{kl} \ln x_{kit} \ln x_{lit} + \sum_{k=1}^{5} \delta_k \ln x_{kit} \ln y^* + \sum_{t=1}^{5} \zeta_t D_t + \xi D_{teach} + v_{it} - u_{it}$$

Table A1: results of the estimation of equation (8) (standard errors into brackets).

parameter	coefficient		parameter	coefficient	
α_0	5.134	(0.788)	β_{24}	-0.133	(0.095)
α_1	0.098	(0.139)	β_{25}	-0.129	(0.181)
α_{11}	-0.005	(0.013)	β_{34}	-0.528	(0.257)
β_1	0.576	(0.202)	β_{35}	-0.375	(0.384)
β_2	-0.845	(0.387)	β_{45}	0.472	(0.202)
β_3	1.839	(0.815)	δ_1	-0.118	(0.038)
β_4	0.699	(0.350)	δ_2	-0.004	(0.036)
β_5	-1.534	(0.688)	δ_3	0.110	(0.085)
β_{11}	-0.040	(0.022)	δ_4	0.019	(0.054)
β_{22}	-0.107	(0.057)	δ_5	-0.122	(0.076)
β_{33}	0.199	(0.311)	ζ_1	0.011	(0.025)
β_{44}	0.148	(0.079)	ζ_2	0.016	(0.027)
β_{55}	0.009	(0.160)	ζ_3	-0.033	(0.033)
β_{12}	0.105	(0.058)	ζ_4	-0.152	(0.038)
β_{13}	-0.141	(0.150)	ζ_5	-0.126	(0.042)
β_{14}	-0.178	(0.087)	Ξ	0.031	(0.045)
β_{15}	0.268	(0.136)	σ^2	0.114	(0.025)

β_{23}	0.489	(0.229)	γ	0.907	(0.023)
$\mathcal{L}$	179.13		η	0.088	(0.017)
OLS $\mathcal{L}$	24.91				

Appendix 2.

Table A2 reports the results of the estimation of equation (15):

$$\ln y_{2it} = \alpha_0 + \alpha_1 \ln y^* + \alpha_{11} (\ln y^*)^2 + \sum_{k=1}^{5} \beta_k \ln x_{kit} +$$
$$+ \sum_{k=1}^{5}\sum_{l=1}^{4} \beta_{kl} \ln x_{kit} \ln x_{lit} + \sum_{k=1}^{5} \delta_k \ln x_{kit} \ln y^* + \rho_0 d + \lambda_1 \ln y^* d + \lambda_{11} (\ln y^*)^2 d +$$
$$+ \sum_{k=1}^{5} \rho_k \ln x_{kit} d + \sum_{k=1}^{5}\sum_{l=1}^{4} \rho_{kl} \ln x_{kit} \ln x_{lit} d + \sum_{k=1}^{5} \zeta_k \ln x_{kit} \ln y^* d + \xi D_{teach} + v_{it} - u_{it}$$

Table A2: Parameters' estimates of equation (15). Standard errors into brackets.

parameter	coefficient		parameter	coefficient	
α_0	6.38	(0.93)	β_{13}	0.18	(0.18)
α_1	0.69	(0.19)	β_{14}	-0.22	(0.09)
α_{11}	-0.05	(0.02)	β_{15}	0.13	(0.16)
β_1	0.29	(0.23)	β_{23}	0.21	(0.25)
β_2	-0.16	(0.47)	β_{24}	0.01	(0.12)
β_3	0.56	(0.91)	β_{25}	-0.08	(0.20)
β_4	0.84	(0.37)	β_{34}	-0.48	(0.28)
β_5	-1.12	(0.70)	β_{35}	-0.19	(0.39)
β_{11}	-0.06	(0.03)	β_{45}	0.34	(0.23)
β_{22}	-0.08	(0.07)	δ_1	-0.12	(0.05)
β_{33}	0.11	(0.31)	δ_2	0.02	(0.05)
β_{44}	0.15	(0.08)	δ_3	-0.07	(0.13)
β_{55}	0.02	(0.17)	δ_4	0.10	(0.07)
β_{12}	0.03	(0.06)	δ_5	-0.10	(0.11)

This part of the table shows the value of the parameters when the dummy is equal to 0 (1994-97)

Table A2: continued

parameter	coefficient		parameter	coefficient	
λ_0	-6.04	(0.20)	ρ_{13}	-0.64	(0.27)
λ_1	-0.42	(0.28)	ρ_{14}	0.08	(0.16)
λ_{11}	-0.02	(0.39)	ρ_{15}	0.10	(0.20)
ρ_1	1.40	(0.23)	ρ_{23}	0.58	(0.31)
ρ_2	-3.02	(0.05)	ρ_{24}	0.03	(0.22)
ρ_3	0.38	(0.05)	ρ_{25}	-0.11	(0.23)
ρ_4	1.83	(0.13)	ρ_{34}	-0.66	(0.51)
ρ_5	0.48	(0.07)	ρ_{35}	0.04	(0.65)
ρ_{11}	0.00	(0.11)	ρ_{45}	0.23	(0.41)
ρ_{22}	-0.26	(1.48)	ζ_1	0.10	(0.08)
ρ_{33}	0.39	(0.27)	ζ_2	0.02	(0.07)
ρ_{44}	0.05	(0.03)	ζ_3	0.23	(0.16)
ρ_{55}	-0.20	(0.59)	ζ_4	-0.19	(0.12)
ρ_{12}	0.32	(0.69)	ζ_5	-0.07	(0.13)
			D_{teach}	0.11	(0.04)
$\mathcal{L}$	218.54		σ^2	0.17	(0.03)
OLS $\mathcal{L}$	49.73		γ	0.95	(0.01)
			η	0.02	(0.01)

This part of the table shows the parameters of the interaction dummy.

Appendix 3.

Analysis of the residuals of equation (15).

The first graph shows the true density of the composed error term $\varepsilon_i = v_i - u_i$

with

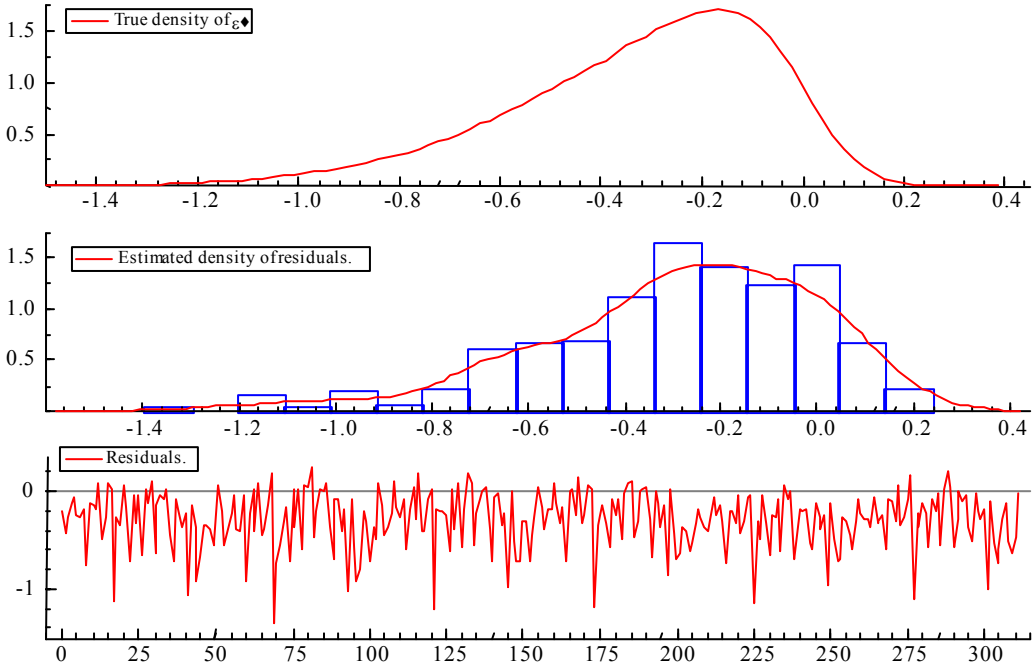
$v_i \sim N(0,\sigma_v^2)$

$u_i = |U_i|$

$U_i \sim N(0,\sigma_u^2)$

The second graph shows the estimated density of the residuals, from the estimation of equation (15). The last graph shows the distribution of the 312 residuals.

The residuals seem to be in broad agreement with their theoretical distribution and there are no serious outliers, indicating that the results are robust.



Appendix 4

Selection of models M1 to M4 estimated in Section 4.

The comparison of models in Section 4 was made by means of LR tests because in each case one model was nested in the other. The restrictions imposed in each case are specified hereinafter.

For all the models:
N=52
T=6
NxT=312
f(.) is the translog distance function, with n = 27 parameters

Model 4: 5 intercept dummy variables

$M4= \alpha_0 + f(.) + \delta_1 D_1 + \delta_2 D_2 + \delta_3 D_3 + \delta_4 D_4 + \delta_5 D_5 + \varepsilon_{it}$

where

D_1=1 for time 2 and 0 else
D_2=1 for time 3 and 0 else
D_3=1 for time 4 and 0 else
D_4=1 for time 5 and 0 else
D_5=1 for time 6 and 0 else

Model 3: quadratic time trend

$M3 = \beta_0 + f(.) + \beta_1 t + \beta_2 t^2 + \varepsilon_{it}$

Model 2: linear time trend

$M2 = \gamma_0 + f(.) + \beta_1 t + \varepsilon_{it}$

Model 1: no time effect

$M1 = \varphi_0 + f(.) + \varepsilon_{it}$

The restriction(s) imposed in the tests were the following:

1) M1 vs M4
   M1 is nested in M4 if
   $\delta_1=\delta_2=\delta_3=\delta_4=\delta_5=0$
   number of restrictions:5

2) M1 vs M3

M1 is nested in M3 if
$\beta_1=\beta_2=0$
number of restrictions:2

3) M1 vs M2
M1 is nested in M2 if
$\beta_1=0$
number of restrictions:1

4) M2 vs M3
M2 is nested in M3 if
$\beta_2=0$
number of restrictions:1

5) M3 vs M4
M3 is nested in M4 if
$\delta_3 = \delta_2 - \delta_1$
$\delta_4 = 2\delta_2 - 3\delta_1$
$\delta_5 = 3\delta_2 - 4\delta_1$
number of restrictions: 3

This comes from observing that M4 could be reparameterised as
$\alpha_0 = \beta_0 + \beta_1 + \beta_2$
$\delta_1 = \beta_1 + 3\beta_2$
$\delta_2 = 2\beta_1 + 8\beta_2$
$\delta_3 = 3\beta_1 + 15\beta_2$
$\delta_4 = 4\beta_1 + 24\beta_2$
$\delta_5 = 5\beta_1 + 35\beta_2$