Data Integration and Productivity Estimation at a Firm Level

Filippo Oropallo, Stefania Rossetti[1]

Abstract Recent years have seen a dramatic surge of interest for the estimation of productivity at firm/plant level both for the implementation of usual performance analysis and for the study of its influence on aggregate patterns. The use of complex data bases, that is longitudinal and cross-section integrated microdata coming from different data sources, improve the possibilities of producing correct estimates at firm level. This paper uses a balanced panel built from a multi source dataset including Business Register, External trade data, Employment data and Financial data for the period 2001-2008. The paper focuses on the estimation of Total Factor Productivity and compares the results of two different estimation techniques.

Keywords: Panel data, Firm performance, Total factor productivity,
Jel codes: C14, C23, D24, L6

[1] Filippo Oropallo, Istat; email: oropallo@istat.it
Stefania Rossetti, Istat; email: stefania.rossetti@istat.it

1 Data description

The longitudinal data base of small and medium-size enterprises includes 76,464 companies, constantly active between 2001 and 2008. Data, for each business unit, are taken from Balance sheets, Business Register, Foreign Trade statistics and from short term statistics on employment and wages. Data linkage of different sources is implemented through an exact matching technique by the identification number provided in the business register and consider the changes in the business unit observed in different years taken from administrative data. The balanced panel dataset includes eight observations for each unit summing up to 611,712 records. The dataset represents 32% of firms with more than 9 employees operating in 2008 and slightly less than 30% of the corresponding employment. However, large enterprises (with at least 500 employees) are not represented in the panel, this is why it has been named small and medium-size enterprises database. The average size of the firm in the panel is 35,4 employees; about 42% of the firm belongs to industry excluding construction and about 46% in services; the remaining firms operate in the construction sector. The dataset contains about 40 basic variables concerning structural characteristics, balance sheet and behavioural information. The variables included allow the computation of a wide number of structural and performance indicators (see Monducci et al. 2010); some of them are reported in table 1.

Table 1: Performance indicators at a firm level – Year 2008 (median values)

Indicators	Median values
Average persons employed	19.0
Labour productivity (000€)	43.4
Labour cost per employee (000€)	34.3
Capital intensity (000€)	28.9
Immaterial capital intensity (000€)	0.8
Share of exported turnover (%)	4.0
Exports per employee (000€)	6.4
Vertical integration (% of value added on turnover)	28.5
Gross profitability (%)	16.9
Return on equity (%)	5.2
Debt ratio (%)	38.9

Source: Our elaboration on Istat and Administrative data

The development and the use of a longitudinal dataset poses at least two main methodological issues:

1) Status change of enterprise (mainly because of mergers and acquisitions). We have excluded firms which have undergone a status change in the period considered, that is we have built a balanced panel, avoiding subjective choices in the treatment of different types of event .
2) Attrition bias: a balanced panel includes only surviving firms. This introduce an obvious distortion problem of self selection type. In order to deal with this problem we have introduced a correction based on a survival function of the firm in the estimate phase.

2 Productivity estimation

Our analysis focuses on the estimate of Total factor productivity (TFP) at firm level, for the subset of manufacturing firms and we experiment two techniques. The first one refers to a theoretical productive model of flexible combination of input factors (labour and capital) and for each firm the distance between its output and the optimal output level is considered the spread of efficiency explained by TFP. We assume a functional form for the production function of the translog type:[2]

$$\ln Y_{it} = \beta_0 + \beta_1 \ln L_{it} + \beta_2 (\ln L_{it})^2 + \beta_3 \ln K_{it} + \beta_4 (\ln K_{it})^2 + \beta_5 \ln L_{it} \ln K_{it} + \tau_t + u_i + \varepsilon_{it} \quad (1)$$

It represents an approximation of an unknown functional form which accounts for substitutability of input factors, namely labour (L), measured by the average number of employees, and capital (K), measured by capital costs (usage of capital). The output (Y) is measured by the value added, that is the value of production diminished by intermediate inputs costs; the introduction of time dummies allows for trend effects and technical progress.

The estimation of such a function with OLS raises a simultaneity bias problem. In fact the firm observes at least part of its own TFP at a point in time early enough to influence its factor input decision. This means that some variables on the right-hand side of equation (1) and the error term are correlated, which makes OLS estimates biased.

To avoid this problem we first use a fixed-effect panel regression, under the assumption that the part of the TFP, that influences the firm behaviour, is time invariant. If this is the case, the residual (or error) term can be split in two parts: one representing the "true" error term and the other (the fixed-effect term) representing the firm specific TFP, that is the spread of efficiency between the firm output level and the optimal output level. In practice, in a fixed-effect panel regression the TFP is given by the following exponential function of the residual u_i:

$$TFP_i^{fe} = \exp(u_i) \quad (2)$$

The results of this estimate are reported in table 2.

A second method frequently used in the related literature to avoid the simultaneity bias problem is the one originally proposed by Olley and Pakes (1996). The method involves two stages and uses firms investments as a proxy of unobserved productivity shocks and produces consistent estimates for the coefficients of labour and capital. An advantage of this method is that it allows for time variability of TFP.

In the first stage, the coefficient of labour input (L) is estimated through a log-linear functional form of value added that includes labour input (L), Capital input (K) and investments[3] (I), inserted in a third order polynomial form:

$$\ln Y_{it} = \beta_0 + \beta_1 \ln L_{it} + \beta_{21} \ln K_{it} + \beta_{22} (\ln K_{it})^2 + \beta_{23} (\ln K_{it})^3 + \beta_{31} \ln I_{it} + \beta_{32} (\ln I_{it})^2 + \beta_{33} (\ln I_{it})^3 + \quad (3)$$
$$+ \beta_{41} \ln K_{it} \ln I_{it} + \beta_{42} (\ln K_{it} \ln I_{it})^2 + \beta_{43} (\ln K_{it} \ln I_{it})^3 + settore_{it} + clad_{it} + \tau_t + \varepsilon_{it}$$

[2] Greene, 1993 e Battese, Coelli, 1995.

[3] Investments are estimated with the following equation: *Inv*(t)=*Imm*(t+1)-*Imm*(t)+*d*(t); where *Imm* represent the stock of capital and d a depreciation factor.

Table 2: Fixed effect estimation of a Translog production function – Years 2001-2008

	coefficiente	standard error	T
ln(L)	1.580	0.0129	122.8
ln(K)	0.153	0.0087	17.5
$(lnL)^2$	0.136	0.0016	85.7
$(lnK)^2$	0.031	0.0005	63.8
ln(L)ln(K)	-0.166	0.0013	-129.0
β_0	6.831	0.0468	146.1
Manufacturing firms			31,904
R^2 within			0.50
R^2 between			0.88
R^2 overall			0.82
corr(u_i, Xb) (a)			0.36
Fraction of var. u_i (b)			0.6

Source: Our elaboration on Istat and Administrative data
(a) Correlation between regressors and the fixed-effect term (latent heterogeneity).
(b)The fixed-effect term is statistically different from zero and it accounts for 60% of unexplained variability of the model.

In the second stage the method also deals with the problem of self-selection (attrition bias) introducing the survival probability of each firm as auxiliary information resulting from the estimation of a survival function which includes employees, capital stock, investments and activity sector. The estimate is obviously performed including in the dataset also the firms that in the period considered have ceased their activity. The average survival rate is equal to 79.1% but varies for each firm according to its characteristics.
Then the estimate includes the following function Φ:

$$\Phi_{it} = (\ln Y_{it} - \hat{\varepsilon}_{it} - \hat{\beta}_1 \ln L_{it}) / survival_i \qquad (4)$$

which is equal to the difference between the log of value added and the residual of the first stage estimate and the contribute of labour input (*L*), divided by the survival rate.
In the second stage the dependent variable *V* is given by the difference between the log of value added and labour input (*L*) multiplied by the estimated coefficient β_1:

$$\ln V_{it} = \alpha_0 + \beta_2 \ln K_{it} + \beta_{31}\Phi_{it-1} + \beta_{32} \ln \Phi^2_{it-1} + \beta_{33} \ln \Phi^3_{it-1} + settore_{it} + clad_{it} + \tau_t + \varepsilon_{it} \quad (5)$$

Since the variable *V* is a non-linear function of Φ, we use a non-linear iterative estimation technique.[4]
Finally we obtain an alternative estimate of TFP equal to the exponential of difference between observed output and estimated output using consistent parameters for input factor:

$$TFP^{op}_{it} = \exp(\ln Y_{it} - (\hat{\beta}_0 + \hat{\beta}_1 \ln L_{it} + \hat{\beta}_2 \ln K_{it} + \varepsilon_{it})) \quad (6)$$

The results of this estimate are reported in table 3.

[4] Arnold, Hussinger 2005.

Table 3- Olley Pakes estimation (a)- First and second stage

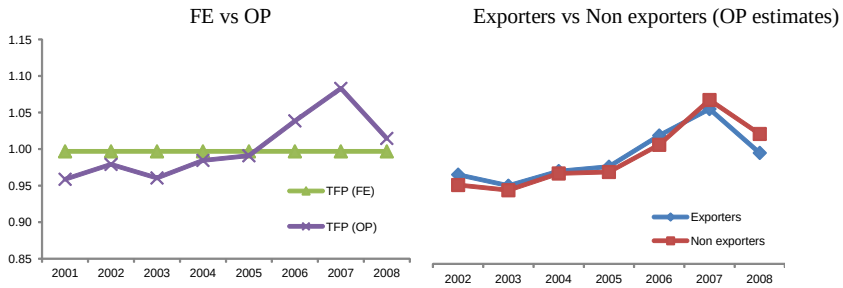
	coefficient	standard error	t
ln(L)	0.853	0.003	340.7
ln(K)	0.241	0.008	29.2
Φ_{t-1}	85.783	2.893	29.7
R^2 adjusted			0.24

Source: Our elaboration on Istat and Administrative data
(a) Dummy variables for activity sector, size class and time are included in the model .

3 Main results

The graphs show the difference between fixed-effect and *OP* estimates taking into consideration the median values. In the first case the method produce a constant estimate for each firm, while the second method show the evolution of TFP, which grows through time and reaches a maximum in 2007. The O-P method allows to evaluate different TFP trend for sub-groups of firms. An example is given in the right–hand side of the graph for exporters and non exporters.

Graph 1 –Total Factor Productivity estimation (median values)



Source: Our elaboration on Istat and Administrative data

References

1. Arnold J. M., Hussinger K.: Export behavior and firm productivity in German manufacturing: a firm-level analysis. Review of World Economics 141 (2): 219-242 (2005)
2. Battese,G . E. and Coelli, T. J.: A Model For Technical Inefficiency Effects in a Stochastic Frontier Production Function for Panel Data, Empirical Economics, 20, pp. 325-332 (1995).
3. Greene, W. H.: The Econometric Approach to Efficiency Analysis, in Fried, H. O., Lovell, C. A. K. and Schmidt, S. S. (eds), The Measurement of Productive Efficiency (Oxford University Press, Oxford 1993)
4. Monducci R., Anitori P., Oropallo F., Pascucci C.: Crisi e ripresa del sistema industriale italiano: tendenze aggregate ed eterogeneità delle imprese. Rivista di Economia e Politica Industriale – Journal of Industrial and Business Economics Vol. 37(3): 93-116 (2010)
5. Olley, S., and A. Pakes: The Dynamics of Productivity in the Telecommunications Equipment Industry. Econometrica 64 (6): 1263–1297 (1996)