THE EFFECT OF UNOBSERVED HETEROGENEITY IN STOCHASTIC FRONTIER ESTIMATION: COMPARISON OF CROSS SECTION AND PANEL WITH SIMULATED DATA FOR THE POSTAL SECTOR

C. Cazals - *TSE (IDEI-GREMAQ)*
P. Dudley - *Royal Mail Group*
J.-P. Florens - *TSE (IDEI-GREMAQ)*
M. Jones - *Royal Mail Group*

1. Introduction

Efficiency analysis : increasingly used in regulated sectors.
→ regulation schemes: based on benchmarking.

Two main approaches:

- non parametric (DEA, FDH, m-frontier, …)
- parametric (stochastic frontier analysis)

Most often applied method: stochastic frontier analysis (SFA)
→ used with cross-section or panel data

In this paper we examine the application of SFA method and assess its estimation of inefficiency when applied to cross section and panel data.

By using ***simulation methods***, we look at the effect of unobserved heterogeneity on the estimates of inefficiency in both cross section and panel.

Result: estimation of inefficiency can be significantly different between cross-section and panel.

Application with actual data from UK postal sector .

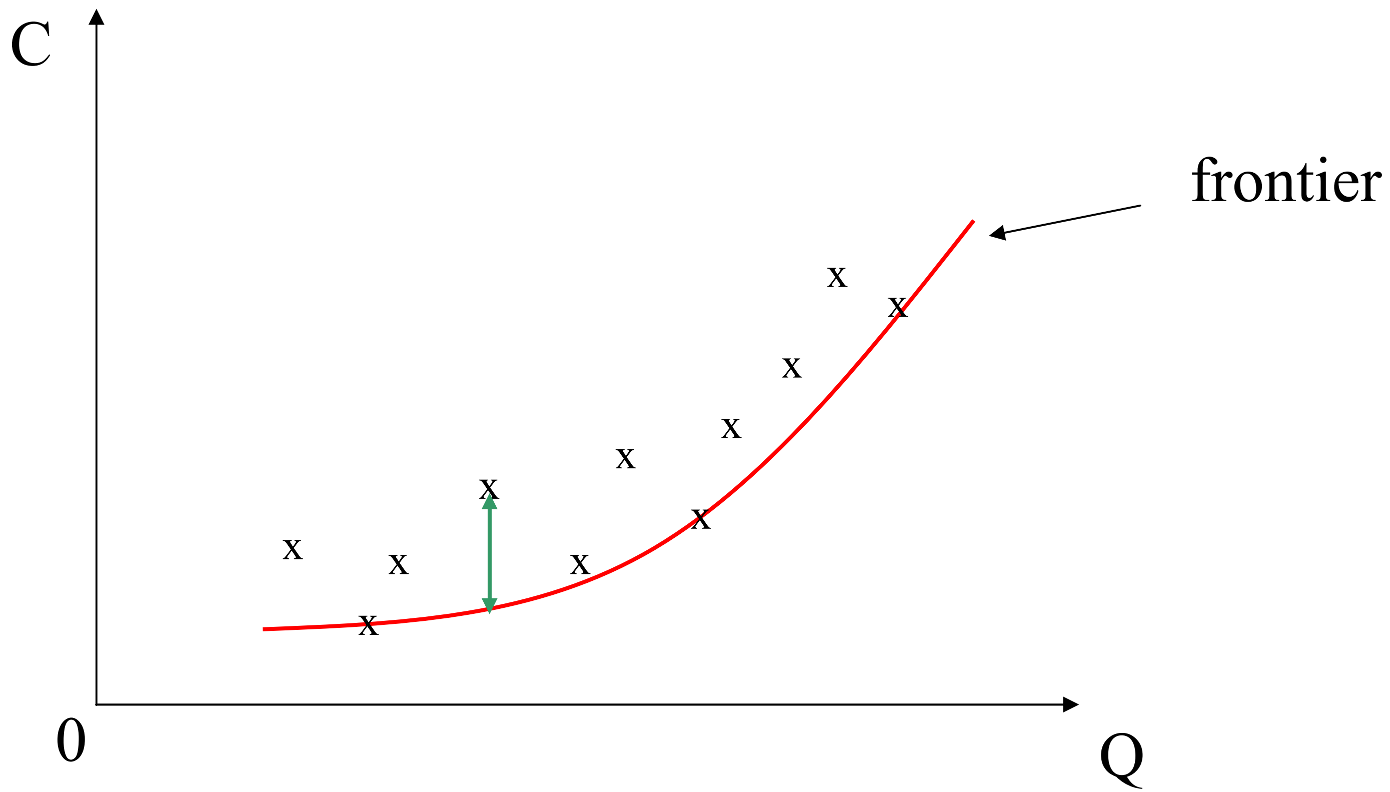
2. Estimation methodology for stochastic frontier analysis

Frontier functions: useful to evaluate performance of production units in relation to the performance of other units, and obtain some efficiency ranking.

Production frontier: searching, for a given level of input, the unit which produces the maximum output.

Cost frontier: searching, for a given level of output, the unit which produces with a minimal cost.

Here we consider cost frontier.


C
0
Q
frontier

Sample of N production units, with information about:

Cost: C
Output: Q
Environmental variables: Z

Application to the delivery process in the postal sector:

C = delivery costs,
Q = delivered mail,
Z = delivery area, number of delivery points, type of delivery zone (rural, urban, …), ….

Stochastic cost frontier model :

$$C = f(Q, Z) + \overbrace{u + \varepsilon}^{v}$$

standard random error term (noise)

inefficiency (≥ 0)

If production units observed for one date: cross-section data

If production units observed for several dates: panel data

❖ ***Cross-section data***: variables indexed by i = 1, …, N.

In most of practical implementation: use of logarithm of variables and Cobb-Douglas (or Translog) functional form for the frontier.

Then the model may be written for example as:

$$C_i = A Q_i^{\beta} Z_i^{\delta} e^{u_i + \varepsilon_i}$$

or, by taking logarithm: $c_i = \alpha + \beta q_i + \delta z_i + u_i + \varepsilon_i$

where c, q and z = Ln of C, Q and Z, and $\alpha = Ln\ A$.

Cost inefficiency: $\dfrac{C_i}{C_i^F} = e^{\varepsilon_i}$

Usual assumptions:

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

$$\varepsilon_i \sim N^+(0, \sigma_\varepsilon^2)$$

u_i and ε_i independent, and u_i independent from q_i and z_i

Most efficient estimation method: maximum likelihood.

Log-likelihood function:

$$l = \text{constant} - NLn\sigma + \sum_{i=1}^{N} Ln\Phi\left(\frac{v_i\lambda}{\sigma}\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{N} v_i^2$$

where $v_i = u_i + \varepsilon_i$, $\sigma^2 = \sigma_u^2 + \sigma_\varepsilon^2$ and $\lambda = \dfrac{\sigma_\varepsilon}{\sigma_u}$

Use of conditional distribution of ε given *v* to estimate the production unit specific inefficiency (Jondrow, Lovell, Materov and Schmidt (1982)).

Estimator of the inefficiency term (Battese and Coelli (1988)):

$$E(e^{\varepsilon_i} \mid v_i) = \left(\frac{1 - \Phi(\sigma_* - \frac{\mu_{*i}}{\sigma_*})}{1 - \Phi(-\frac{\mu_{*i}}{\sigma_*})} \right) \exp(-\mu_{*i} + 0.5\sigma_*^2)$$

where $\mu_{*i} = \frac{v_i \sigma_\varepsilon^2}{\sigma^2}$ and $\sigma_* = \frac{\sigma_\varepsilon^2 \sigma_u^2}{\sigma^2}$

- ***Panel data***: variables indexed by i = 1, …, N, and t= 1,…, T.

Model is written, in the case of time-invariant inefficiency:

$$c_{it} = \alpha + \beta q_{it} + \delta z_{it} + u_{it} + \varepsilon_i$$

Assumptions about *u* and *ε*: similar to cross-sectional model.

Two approaches in standard applications:

- ✓ fixed effects
- ✓ random effects (more often applied)

Estimation of RE models by maximum likelihood method.

Log-likelihood function:

$$l = const - \frac{N(T-1)}{2} Ln\sigma_u^2 - \frac{N}{2} Ln(\sigma_u^2 + T\sigma_\varepsilon^2) +$$

$$\sum_i Ln\left[1 - \Phi(-\frac{\mu_{*i}}{\sigma_*})\right] - \left(\frac{v'v}{2\sigma_u^2}\right) + \frac{1}{2}\sum_i \left(\frac{\mu_{*i}}{\sigma_*}\right)^2$$

where $\mu_{*_i} = \frac{T\sigma_\varepsilon^2 \overline{v}_i}{\left(\sigma_u^2 + T\sigma_\varepsilon^2\right)}$ and $\sigma_*^2 = \frac{\sigma_u \sigma_\varepsilon}{\left(\sigma_u^2 + T\sigma_\varepsilon^2\right)}$

Estimator of the inefficiency term:

$$E\left[\mathrm{e}^{\varepsilon_i} \mid v_i\right] = \frac{1-\Phi\left[\sigma_* - (\mu_{*_i} / \sigma_*)\right]}{1-\Phi\left[\mu_{*_i} / \sigma_*\right]} \exp\left(-\mu_{*_i} + 0.5\sigma_*^2\right)$$

Drawback of these standard panel models: if there exists some persistent unobserved heterogeneity, it will be considered as inefficiency.

Here: examination of the magnitude of the difference between inefficiency scores in cross sectional and panel models, due to the presence of unobserved heterogeneity.

3. Analysis of the effect of unobserved heterogeneity with a simulated model

Simulated data sets used to examine the effect of the presence of unobserved heterogeneity on the estimation of inefficiency in cross-section and panel models.

inefficiency

Data generating process: $c_{it} = \alpha + \beta q_{it} + u_{it} + \varepsilon_i$

We assume here: $u_{it} = w_{it} + v_i$,

where v_i : ***unobserved heterogeneity***, assumed $N(0, \sigma_v^2)$

w_{it} : statistical noise, assumed $N(0, \sigma_w^2)$

and ε_i assumed $N^+(0, \sigma_\varepsilon^2)$

Chosen values for parameters in the simulation exercise:

- ✓ q_{it} generated from the model: $q_{it} = 0.18 + 0.94 q_{i,t-1} + \zeta_{it}$

$N(0;0.0225)$

- ✓ α= -3.7 and β = 0.94
- ✓ $\sigma^2 = \sigma_w^2 + \sigma_v^2 + \sigma_\varepsilon^2 = 0.05$
- ✓ Let us define: $\gamma = \dfrac{\sigma_\varepsilon^2}{\sigma^2}$ and $\lambda = \dfrac{\sigma_v^2}{\sigma_w^2 + \sigma_v^2}$

We consider 2 different values for γ, $\gamma = 0.5$ and $\gamma = 0.9$

and 3 different values for λ: 0.1, 0.5 and 0.9.

For each value of $(\gamma,\lambda) \rightarrow$ 50 samples with a size N=500 are generated.
2 time periods considered.

For each sample: estimation of a panel model, and cross-section models for t=1 and t=2.
$\rightarrow$ use of ML method (RE model for panel data).

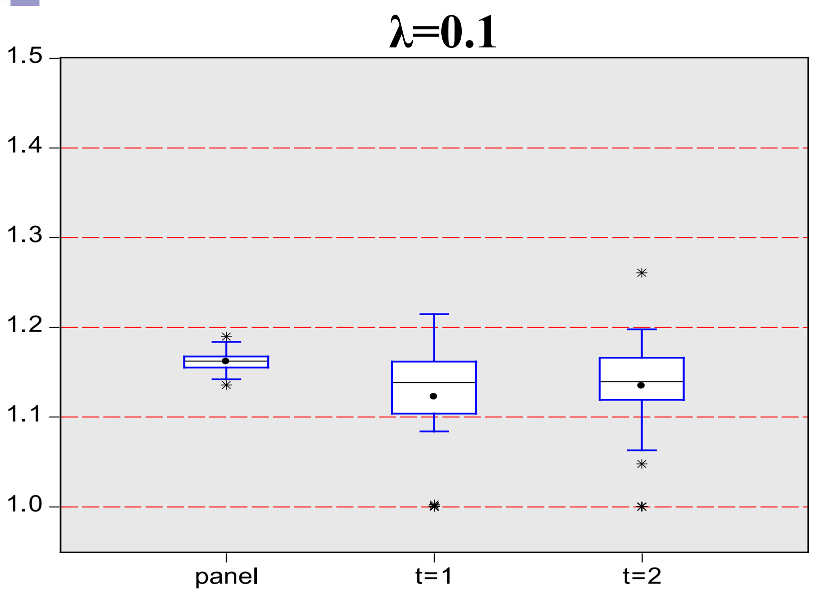
Estimation of inefficiency scores $\rightarrow$ values ≥ 1.

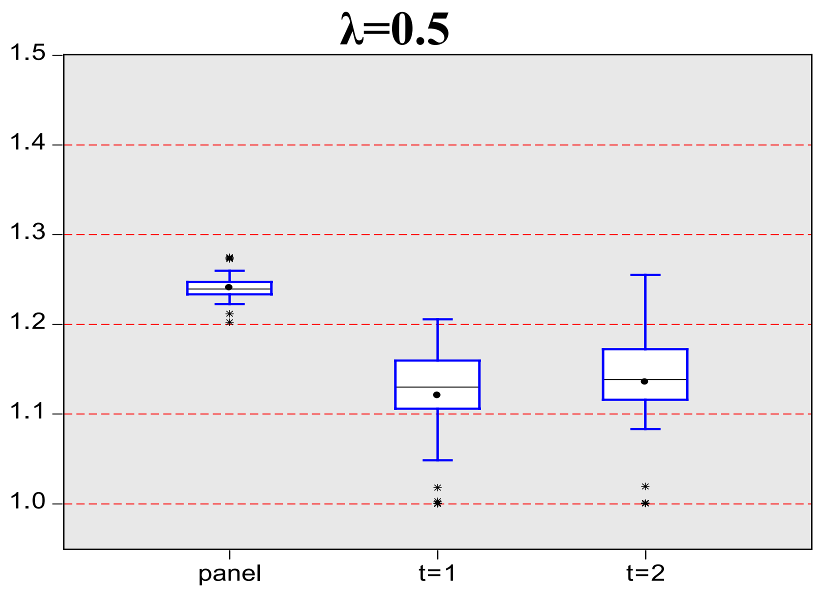
Mean value for the "true" inefficiency :

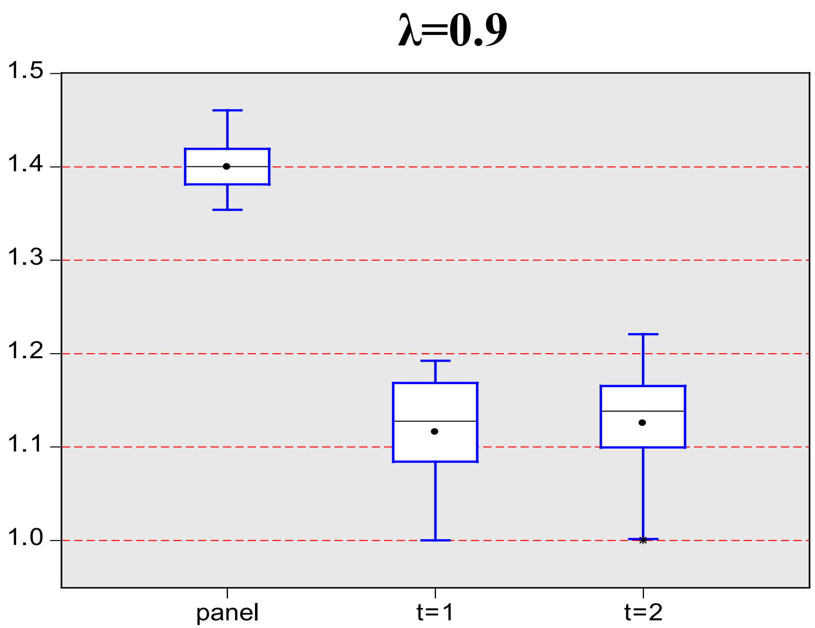
- 1.134 when $\gamma = 0.5$
- 1.184 when $\gamma = 0.9$.

Comparison of means of estimated inefficiency scores obtained with panel to those obtained with cross-section over the 50 samples.

→ ***Box plots*** for the means of inefficiencies

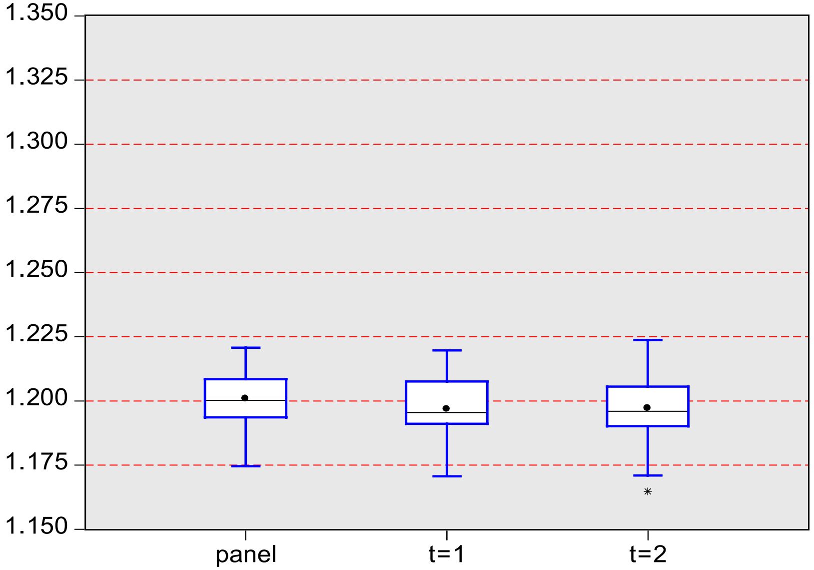




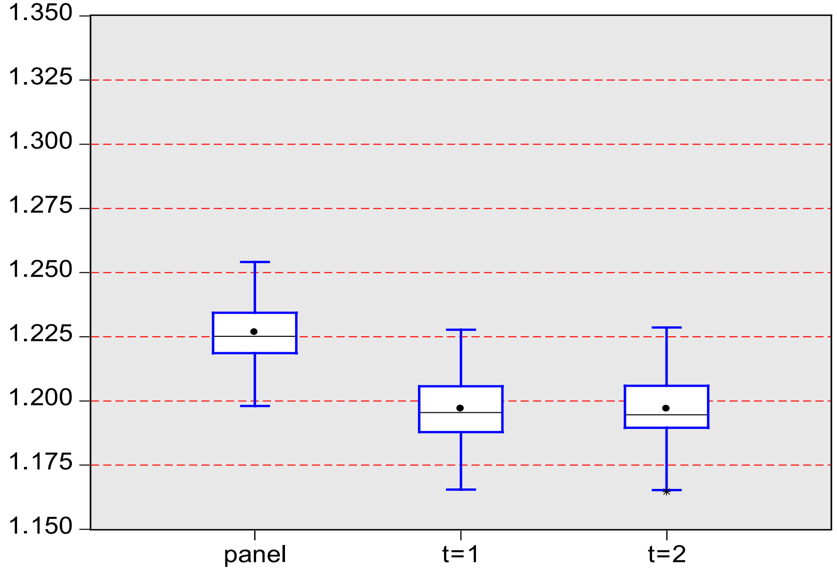


Case 1: γ=0.5

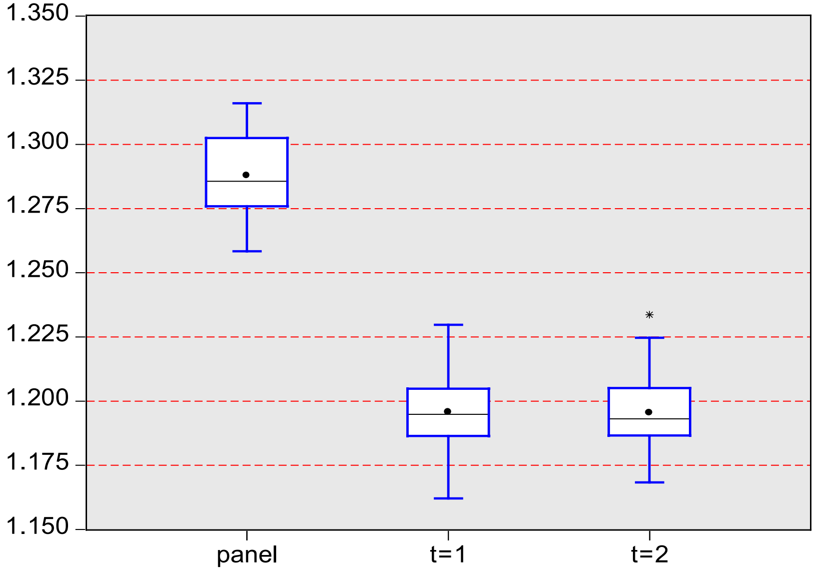
λ=0.1



λ=0.5



λ=0.9



Case 2: γ=0.9

Quartiles for means of inefficiency scores

$\gamma = 0.5$

	●=0.1			●=0.5			●=0.9		
	panel	t=1	t=2	panel	t=1	t=2	panel	t=1	t=2
p25	1.155	1.102	1.119	1.233	1.097	1.113	1.381	1.080	1.098
p50	1.162	1.138	1.140	1.238	1.128	1.138	1.400	1.112	1.135
p75	1.167	1.161	1.167	1.246	1.159	1.178	1.419	1.168	1.163
IQR	*0.012*	*0.059*	*0.048*	*0.013*	*0.062*	*0.065*	*0.038*	*0.088*	*0.065*

$\gamma = 0.9$

	●=0.1			●=0.5			●=0.9		
	panel	t=1	t=2	panel	t=1	t=2	panel	t=1	t=2
p25	1.193	1.190	1.190	1.218	1.187	1.188	1.275	1.185	1.185
p50	1.200	1.195	1.195	1.224	1.194	1.194	1.285	1.193	1.193
p75	1.208	1.207	1.205	1.234	1.205	1.205	1.301	1.204	1.205
IQR	*0.015*	*0017*	*0.015*	*0.016*	*0.018*	*0.017*	*0.026*	*0.019*	*0.020*

Main results:

- mainly when $\gamma = 0.5$: distributions for inefficiency are broader in the cross-section cases than in the panel cases.
- inefficiency : better estimated in the cross section case when the part of the variance due to unobserved heterogeneity increases.

Main comment:

Panel estimation methodology tends to attribute more of the unobserved heterogeneity component to the inefficiency component.

Results from the panel and cross section analyses: upper and lower bounds on the inefficiency estimate

3. Application of SFA to real data: for delivery activity in the postal sector

Data set for 1334 delivery offices observed for 6 years between 2003/04 and 2008/09.

Observed variables:

- number of worked hours (C),
- volume of delivered mail (Q),
- number of delivery point (DP),
- surface of the delivery zone (AR),
- proportion of business delivery points (prop. bus),
- indicator of the type of delivery area (urban, suburb or rural)

Model

$$LnC = \alpha + \beta Ln(Q / DP) + \delta LnZ + u + \varepsilon$$

where Z=(DP, AR/DP, prop. bus, types of delivery area) .

Estimated for cross-section and panel.

Results for cross-section SFA (year 2003/04).

Variables	Coef.	Std. Err.	t-Student
Ln Q/DP	**0.607**	*0.018*	32.82
Ln DP	**1.041**	*0.006*	165.58
Ln AR/DP	**0.057**	*0.005*	10.13
prop. Bus	**1.302**	*0.137*	9.48
urban	**0.026**	*0.029*	0.91
suburb	**-0.028**	*0.029*	-0.97
rural	**-0.051**	*0.033*	-1.53
c	**-2.229**	*0.143*	-15.53
$\gamma = \sigma_{\varepsilon}^{2} / \sigma^{2}$	**0.625**		

Inefficiency scores

mean	*1.122*
st. dev.	*0.066*

Other years: very similar results

Results for panel SFA

Variables	Coef.	*Std. Err.*	t-Student
Ln Q/DP	**0.345**	*0.006*	50.16
Ln DP	**0.932**	*0.007*	127.22
Ln AR/DP	**0.032**	*0.003*	9.74
prop. bus	**1.163**	*0.070*	16.59
urban	**-0.006**	*0.004*	-1.47
suburb	**-0.022**	*0.005*	-3.78
rural	**-0.007**	*0.009*	-0.78
year	**-0.009**	*0.0007*	-13.96
c	**0.308**	*0.089*	3.44
$\gamma = \sigma_{\varepsilon}^{2} / \sigma^{2}$	***0.961***		

Inefficiency scores

mean	*1.324*
st. dev.	*0.216*

Comparison with the simulation exercice:

→ suggest an important unobserved heterogeneity component.

→ close to the case λ=0.9, with γ probably approaching to 0.9.

4. Conclusion:

Standard panel SFA tends to consider unobserved heterogeneity as inefficiency.

When the model is not correctly specified: cross-section method preferable.

With a correctly specified model: better results with panel method (always better to have more information).

Comparison of results in cross-section models and in panel models: useful as a specification test of the model.

Use of information contained in the difference between the two remains : open question for future research