

A propensity score matching method for the link
between accessibility and productivity

ERSA Congress, Jyväskylä

Tom Petersen
Systems Analysis and Economics,
Dept. of Infrastructure, KTH, Stockholm, Sweden
tomp@infra.kth.se

August 27–30, 2003

1 Aim of the study

- Investigate the effect of accessibility on the productivity/efficiency of individual firms in the Øresund region.
- Make predictions about the future effect of a major infrastructure investment: the Øresund bridge
- Effects should be easier to detect the larger the investment

2 Why should there be an effect?

- better access and bigger choice of workers and suppliers;
- higher flexibility;
- access to a bigger market;
- more competition stimulates better efficiency

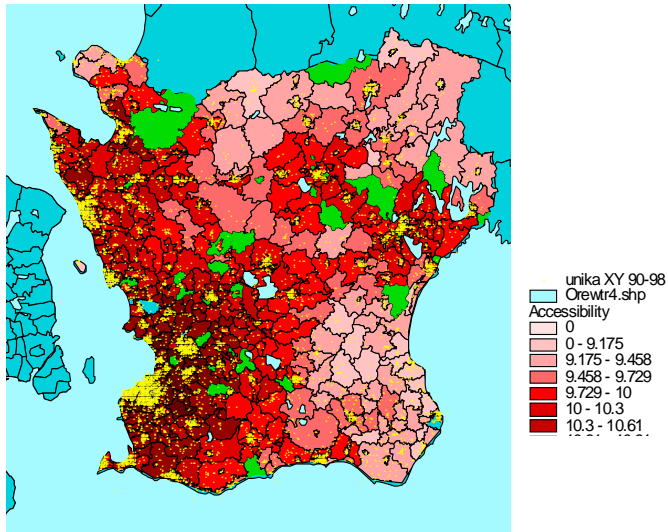
3 Why should there not be an effect?

- firms have already made the best use of their comparative advantages, including location and accessibility (this effect is partly taken into account in the following approach)

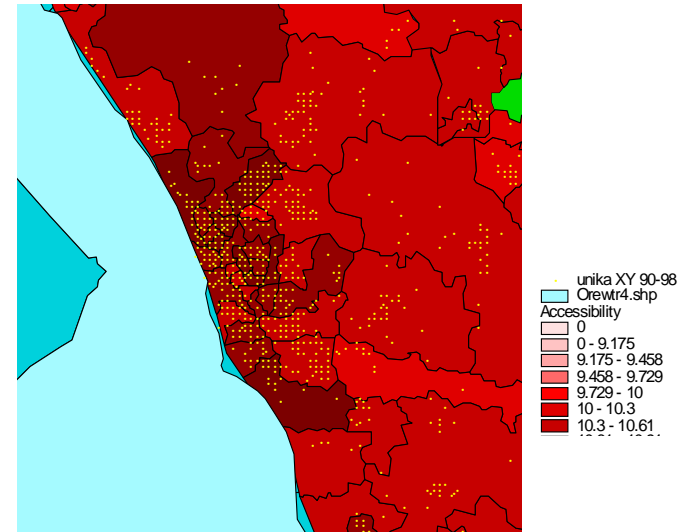
4 Methods:

- Panel data estimation of translog cost function; fixed and random effects
- Propensity score matching method (PSM), using Total Factor Productivity (TFP), Törnqvist index of prod. change, and Fixed effects (FE)
- The propensity score is given by a binary logit model of location in area with "high" or "low" accessibility.

5 The study area



Scania, Øresund region.



Detail: Helsingborg.

6 Parametric vs. non-parametric approach

- Cost function estimation aims at the determination of a parameter, telling the amount of influence of accessibility on the cost. The result depends on the specification, functional form, and the distribution of the residuals.
- In contrast, PSM is a non-parametric method. A non-parametric approach is
 - insensitive to distributional assumptions
 - insensitive to different kinds of misspecification, functional form etc.
 - but there is only a yes/no answer.

7 Effect of treatment

Let

- $D = 1$ for treated, $D = 0$ for non-treated
- Y_1 is outcome for treated, Y_0 outcome for non-treated

We want to estimate the average treatment effect

$$E((Y_1 - Y_0|D = 1, X) \approx E((Y_1|D = 1, X) - (Y_0|D = 0, X)),$$

where all expectations are conditional on background variables X .

8 Test setup

- Our "treatment" = high accessibility (above median in the sample)
- Our outcome = TFP/change in TFP/cost efficiency (i.e. fixed effects from the cost function estimation)
- Hypothesis to be tested: "High accessibility increases productivity/efficiency for individual firms."
- Dataset: 24,630 firms in 7,629 locations, in 24 branch aggregates from the years 1990–98.

9 Why condition on X ?

- firms have different probability to "participate", i.e. to belong to the treatment group
- these differences give rise to selection bias of estimates

Solution 1: Compare only firms with the same values on the background variables.

However,

- X is multidimensional and comparisons are impractical

9.1 Solution 2: Construct the participation probabilities – the “propensity score”

- estimate a binary logit model for participation
- maximise the number of correctly predicted observations
- use predicted probabilities as the propensity score $P(X)$

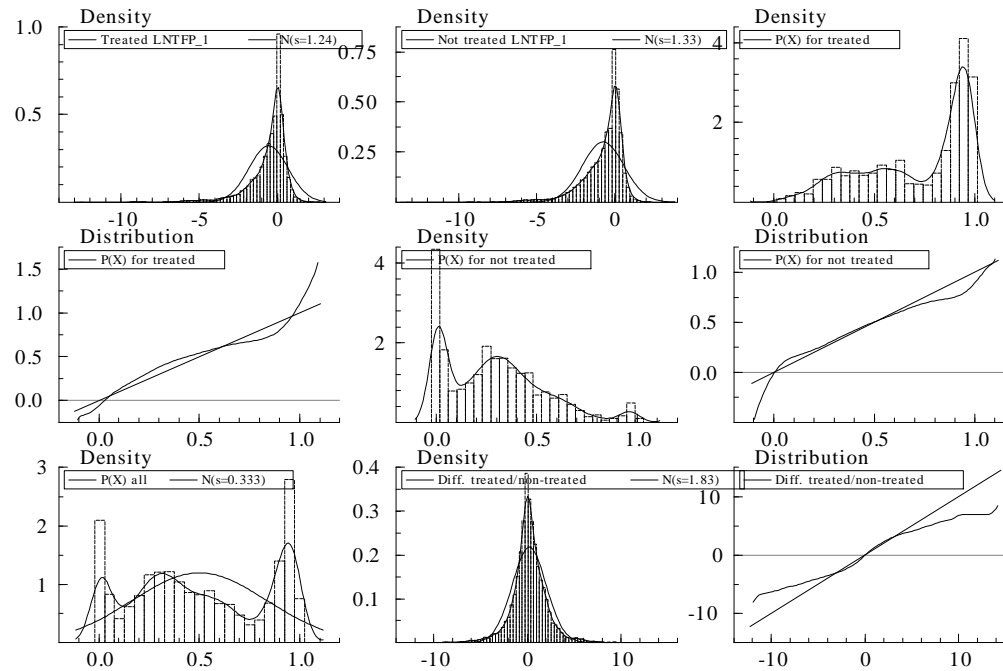


Figure 1: Densities of Y_1 , Y_0 , $(P(X)|D = 1)$, $(P(X)|D = 0)$: the service sector.

10 Matching

Every observation of Y_{1i} is compared with a weighted mean of Y_{0j} observations in the vicinity of $P(X)_{Y_{1i}}$:

$$E((Y_1 - Y_0 | D = 1, P(X))) = \frac{1}{N_1} \sum_{i \in I_1} [Y_{1i} - \sum_{j \in I_0} W_{N_0, N_1}(i, j) Y_{0j}],$$

where

I_a is the set of indices for Y_a , $a \in \{0, 1\}$;

$W_{N_0, N_1}(i, j)$ is a positive valued weight function satisfying $\sum_{j \in I_0} W_{N_0, N_1}(i, j) = 1$ for each i , and

N_a is the number of individuals in I_a , $a \in \{0, 1\}$ (for standardization of weight).

11 Kernels

A kernel is a piecewise continuous function, symmetric around 0 and integrating to 1:

$$K(u) = K(-u), \quad \int_{-\infty}^{\infty} K(u)du = 1$$

We impose bounded support on $[-1, 1]$: $K(u) = 0$ for $|u| \geq 1$. It follows that $K(u)$ has its maximum at $u = 0$.

Example: biweight (quartic) kernel

$$K(u) = \frac{15}{16} (1 - u^2)^2 \cdot I(|u| \leq 1),$$

where $I(\cdot)$ is the indicator function taking the value 1 if the event is true, and 0 otherwise.

12 Weight function

The kernel-based weight function

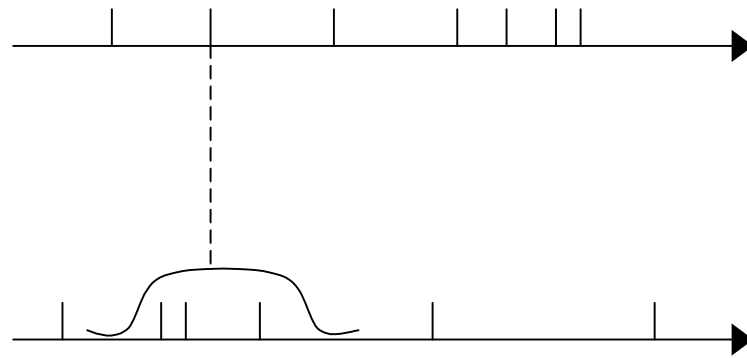
$$W_{N_0, N_1}(i, j) = \frac{K\left(\frac{P(X_i) - P(X_j)}{h}\right)}{\sum_{k=1}^{N_0} K\left(\frac{P(X_i) - P(X_k)}{h}\right)}$$

weights observations higher the closer they are.

For observations that deviate more than the bandwidth h , the weight is 0.

h determines the smoothness of the distribution of differences.

$P(X)$, treated



$P(X)$, non-treated

Figure 2: The matching procedure.

13 Variables for PSM

13.1 Productivity

Total factor productivity, TFP:

$$TFP = \frac{y}{\sum_i s_i x_i},$$

y is output volume; s_i are cost shares for capital, labour and material; x_i are volumes of factor inputs. In logs,

$$\ln(TFP) = \ln(y) - \sum_i s_i \ln(x_i)$$

13.2 Accessibility

Measured by logsums,

$$A_i^n = \mu^{-1} \ln \sum_{j \in L} e^{\mu(v_j^n - c_{ij}^n)},$$

n is individual; i and j are zone indices; L is the total set of zones,

v_j^n is the attractiveness of zone j for individual n ,

c_{ij}^n is the cost of travel between zones, and μ is a scale parameter.

14 Results

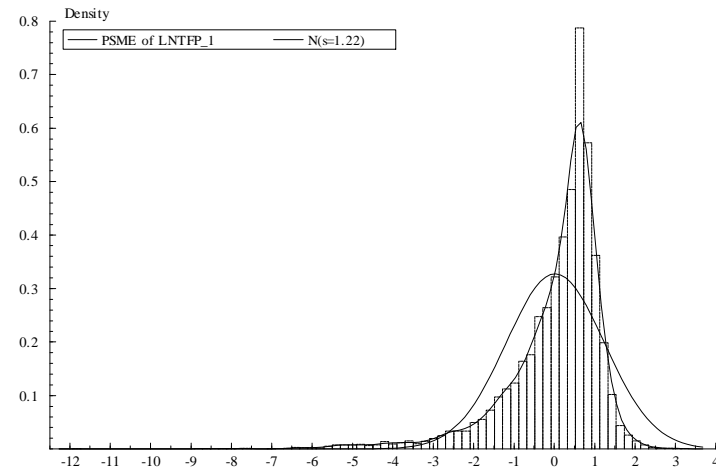


Figure 3: Propensity score matching estimator for the pooled dataset. The mean is 0.023.

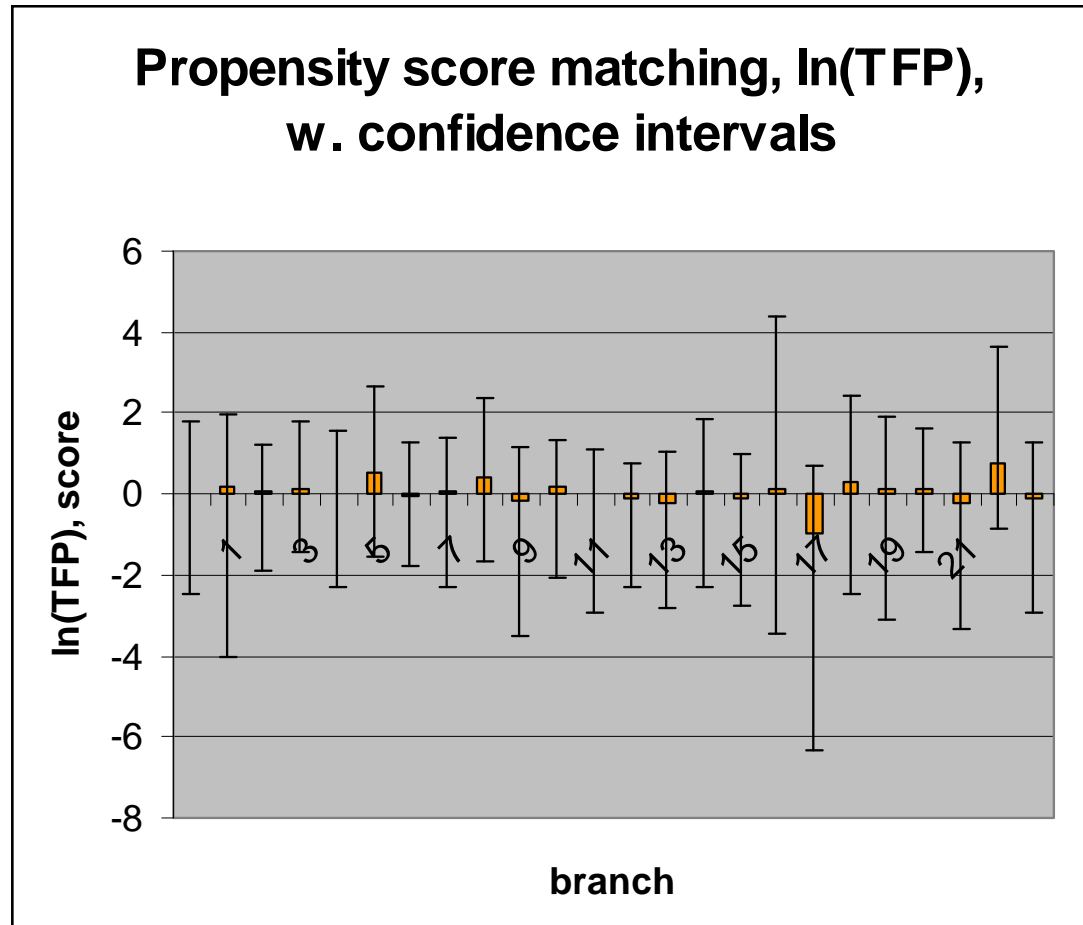


Figure 4: Result of test: no significant effect in any branch.

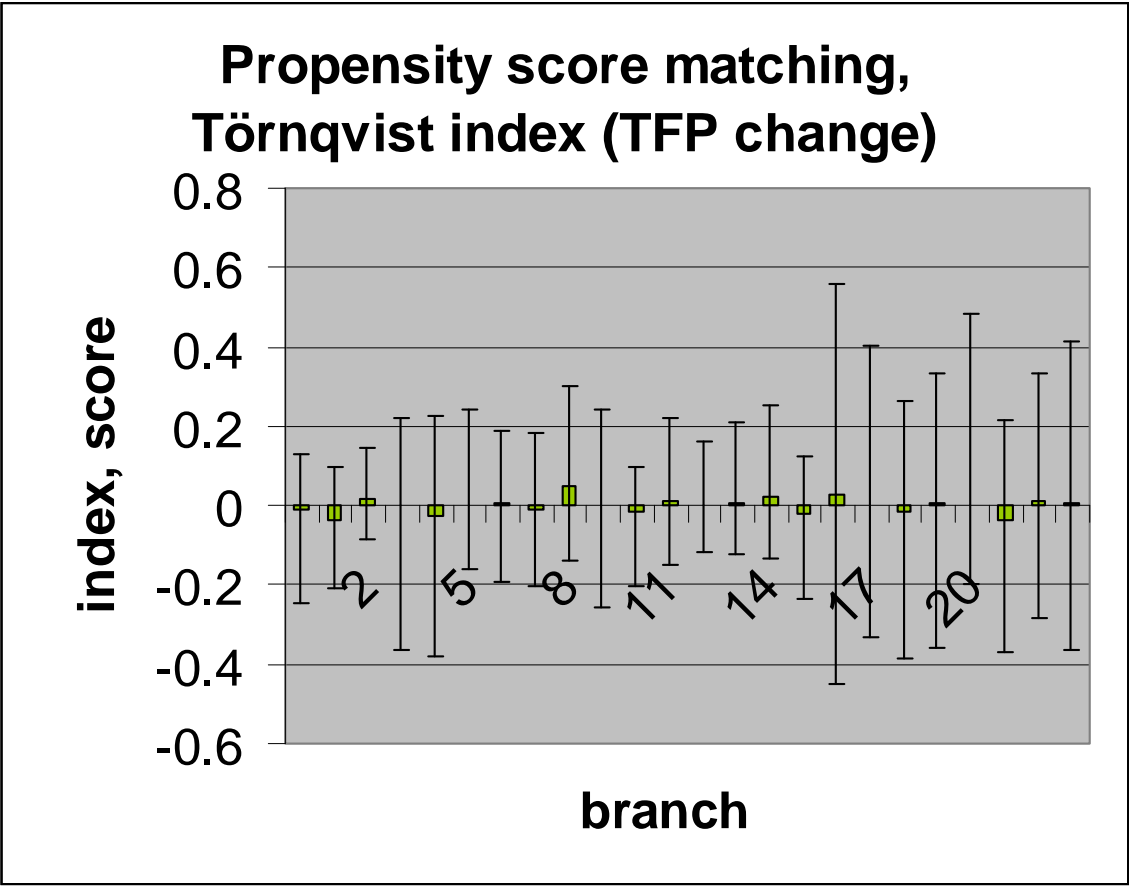


Figure 5: Result for TFP change.

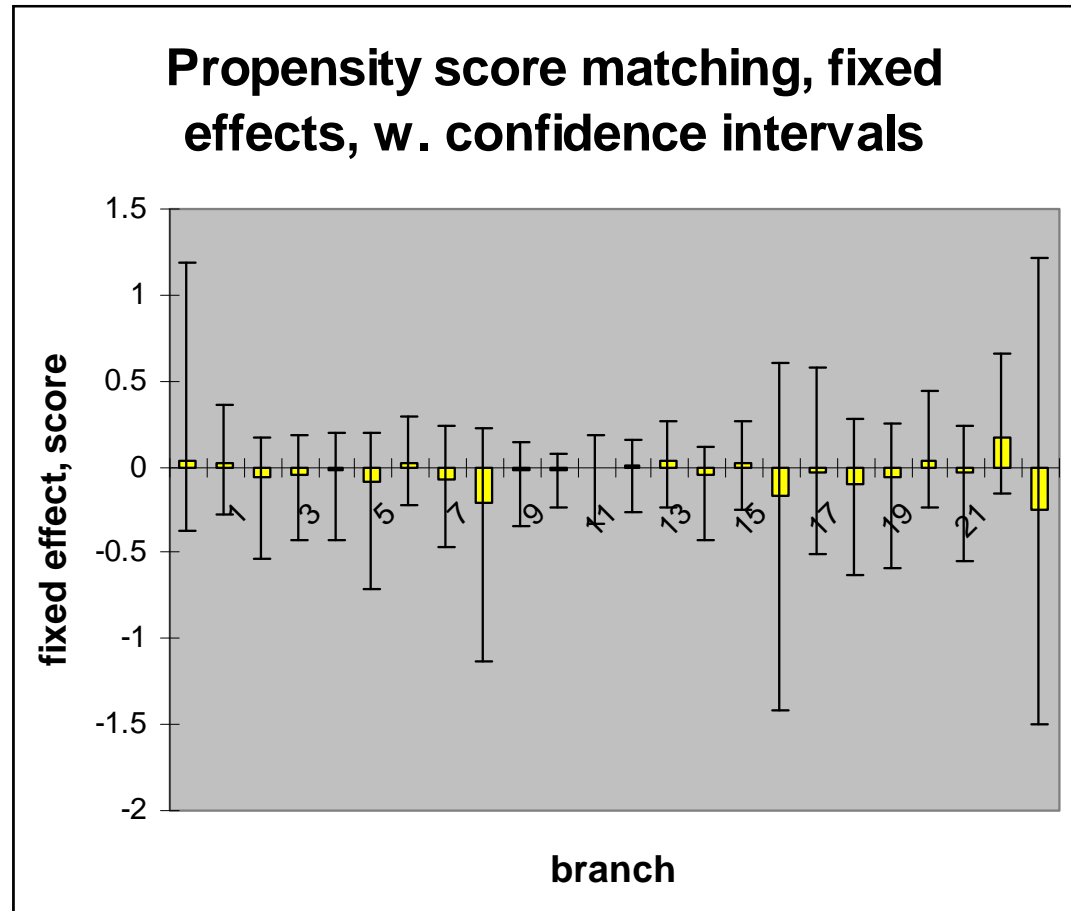


Figure 6: The result for fixed effects (cost efficiency).

15 Propensity score (logit model)

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	75978.7678	0.213	0.285
...
21	59927.3549	0.383	0.510

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	15918.63805	1	0
	Block	15918.63805	1	0
	Model	15918.63805	1	0
...
Step 21	Step	6.13223551	1	0.013
	Block	31970.05103	84	0
	Model	31970.05103	84	0

Classification Table (a)

		Predicted		
		HIGHT		Percentage Correct
Observed		0	1	
Step 1	HIGHT	0	31539 1578	95.24
		1	18327 14846	44.75
Overall Percentage				69.97
...
Step 16	HIGHT	0	27954 5163	84.41
		1	9299 23874	71.97
Overall Percentage				78.18
...
Step 21	HIGHT	0	27853 5264	84.10
		1	9344 23829	71.83
Overall Percentage				77.96

(a) The cut value is .500

Fit statistics of logit model.

78 % correct predictions.

		B	S.E.	Wald	df	Sig.	Exp(B)
AGKAT				49.187	9	0	
OMSATTN		0	0	27.081	1	0	1
JUSTEK		0	0	15.559	1	0	1
NETINVST		0	0	3.985	1	0.046	1
SNI2NR				664.315	56	0	
ANTARBPL	(L)	-0.001	0	1426.975	1	0	0.999
BUPDENS	(L)	1.866	0.046	1671.513	1	0	6.462
DAGBEF_T	(L)	0.001	0	2486.514	1	0	1.001
STORTUNI	(L)	0	0	284.113	1	0	1
LITETUNI	(L)	7.484	1.789	17.493	1	0	1778.776
REGIONSJ	(L)	0.006	0.004	1.755	1	0.185	1.006
LANSDELS	(L)	-0.03	0.006	25.764	1	0	0.97
STORREKO(1)	(L)	-1.799	0.107	281.573	1	0	0.165
STORMARK(1)	(L)	-1.047	0.059	314.763	1	0	0.351
STORREMA(1)	(L)	6.041	1.699	12.637	1	0	420.202
TURISTOM(1)	(L)	2.595	0.081	1033.252	1	0	13.391
TURISTPU(1)	(L)	-0.124	0.05	6.131	1	0.013	0.883
MARKTAX	(L)	0	0	59.188	1	0	1
ÖVRIGTTA	(L)	0	0	804.66	1	0	1
C_ORTKOM(1)	(L)	-0.133	0.027	24.424	1	0	0.875
C_ORTLAN(1)	(L)	-3.471	0.044	6086.383	1	0	0.031
Constant		-3.979	1.711	5.408	1	0.02	0.019

Figure 7: Variables included in the logit model. (L) stands for location specific, in contrast to individual specific.

16 Sources of error

- accessibility is measured from household to firms, not the other way around
- the resolution of accessibility is coarse as compared to the location of firms
- only firms with a unique location in the dataset

17 Conclusions

- no effect of accessibility in this cross-sectional study
- A small cost reduction (4–6% per unit logsum) detected in Construction and Transport sectors, for moving firms during the period 1990–98
- dynamic effects can not be excluded, wait for the after-study!