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

Paper for the 43rd European Congress of the Regional Science Association in Jyväskylä, August 27–30, 2003

Tom Petersen
Systems Analysis and Economics,
Dept. of Infrastructure,
Royal Institute of Technology (KTH),
Teknikringen 78 B, 1st floor
SE-100 44 Stockholm, Sweden
Tel: +46 8 790 96 32
Fax: +46 8 790 70 02
Email: tomp@infra.kth.se

June 15, 2003

Abstract

We present a nonparametric approach to the link between accessibility and productivity, using two methods: propensity score matching and weighted means. The firms located in places with "high accessibility" are regarded as the "treatment group", and the hypothesis tested is whether there is any significant treatment effect. The accessibility is a cross-section for the year 1997. We use a panel dataset consisting of 24,915 individual firms during 1990–98, subdivided by branch. Each firm is geocoded with a 250 m resolution. The study region is Scania, the Swedish part of the new Øresund region. The results show no effect of the "high accessibility" treatment in this set-up; neither on any branch level, nor on the pooled dataset.

Key words: propensity score matching, Oresund, productivity, accessibility.

1 Introduction

The link between infrastructure and productivity has long been under debate, ever since Aschauer (1989) and Munnell (1990). It has also experienced a considerable scrutiny from an econometric point of view, involving panel data (Holtz-Eakin 1994, Chandra & Thompson 2000, Baltagi, Song & Jung 2001) and spatial effects (Moreno, Artis, López-Bazo & Suriñach 1997), not to mention a host of other contributions. In all these studies an aggregate, macroeconomic approach has been utilized. Also in the panel data approaches, the smallest units of observation are county or state. The present paper distinguishes itself not only by the use of data on a microeconomic level, but also by the use of a non-parametric approach, the propensity score matching method.

The paper starts off with a description of the data used in the next section, followed by a section (3) about the two methods, propensity score matching and the weighted means method. The required variables are described in section 4, the results are presented and discussed in sections 5 and 6, and the paper concludes in section 7. An appendix is also included at the end.

2 Data

The analysis is based on account balance data for 24,915 firms for the years 1990–98 in the region of Skåne (Scania) in southern Sweden. The data is collected by Statistics Sweden. Because of the spatial dimension of the problem, and the fact that the economic variables are on firm level, the survey is restricted to firms with only one production unit. However, firms with other production units outside Scania have been included, if they had more than 50 % of their work force in Scania. The data comes from two databases at Statistics Sweden: CFAR (a database on businesses and workplaces) and the Structural Business Statistics database (Företagsregistret, SBS). In SBS, there are data on the income statements and balance sheets of the businesses, but no connection to geography, while in CFAR, there are addresses of each local production unit. Some of the observations are not possible to geocode, other observations disappear in the estimation process because of negative or zero values of the capital and labour variables. The final dataset contains 24,630 firms, with 1–9 observations for each firm. The mean number of observations is about 3.5, and the majority are from the later 4 years (1995–98). In this study, the observations on each firm are averaged over

time in order to obtain a one-to-one relationship between productivity and location. However, the time average has been performed as the last step, i.e. first the relevant variables have been calculated on a year-by-year level, then the variables have been time averaged.

3 Method

3.1 Propensity score matching

We want to estimate a treatment effect on the total factor productivity of individual firms. In our case, the treatment is "high accessibility" or accessibility above the median in Scania, as measured by SAMPERS logsums for worktrips. We denote this $D_i = 1$ if firm i belongs to this group, and $D_i = 0$ otherwise. But firms do not have equal probability to be located just anywhere. Variables like size, industry, ownership, history etc. affect the choice of location. We denote these background variables x_i. Of course, the management of the firm makes an economic judgement from time to another whether it is in the right location, or if it would be better off somewhere else. Different firms derive profit in different ways, some are material intensive and others information intensive. All have different needs, and consider accessibility and value it differently. What we need to do is to control for those background variables x_i and the selectivity bias emerging from the assumption that the firms are already are located in a place optimal for them. The role of the propensity score is to relax the spatial connection of the firm and make it comparable with other firms with different bundles of preferences.

The propensity score method was introduced by Rosenbaum & Rubin (1983, 1985). It accounts for selectivity bias and the differences in subjects receiving a treatment by estimating the probability to receive treatment, given these background variables. This probability, $\Pr(D_i = 1|x_i)$, is called the propensity score. The subjects with similar participation probabilities are grouped together so that the outcome is conditionally independent of whether the subject received treatment or not, or

$$(Y_0, Y_1) \perp D \mid X$$

where Y is the outcome, D the treatment, $D \in \{0, 1\}$ and X background variables. We want to estimate the average treatment effect on the treated, or

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

where we approximate the counterfactual no-treatment effect on the treated group $E(Y_0|D = 1, X)$ with the average outcome of the self-selected no-treatment group $(Y_0|D = 0, X)$ (Heckman, Ichimura & Todd 1998). Instead of conditioning on all the pre-treatment variables X, we introduce the propensity score $P(X) = \Pr(D = 1|X)$, i.e. the probability to participate in the treatment group. To this end, we estimate a binary logit model (see section 4.3) for the alternatives "high accessibility" (HIGH) and "low accessibility" (LOW). We then use the predicted probabilities to be in group HIGH as our propensity score:

$$E((Y_1 - Y_0|D = 1, P(X))$$

3.1.1 Kernel matching

The treatment effect is estimated by contrasting the outcomes of the treatment group (denoted Y_1, indexed by I_1) to the outcomes of a comparable group of non-participants (denoted Y_0, indexed by I_0), by means of the following formula (Heckman et al. 1998, somewhat modified):

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

where $W_{N_0,N_1}(i,j)$ is a positive valued weight function which for each i satisfies $\sum_{j \in I_0} W_{N_0,N_1}(i,j) = 1$, and N_0 and N_1 are the number of individuals in I_0 and I_1, respectively. $W_{N_0,N_1}(i,j)$ is close to one if the distance between individual i and individual j is close in terms of the propensity score, and the weighted sum of Y_{0j} is close to one if there are many such comparable individuals in the neighbourhood.

The weight function $W_{N_0,N_1}(i,j)$ is implemented by means of a kernel, which is a piecewise continuous function, symmetric around zero and integrating to one (Härdle & Linton 1994):

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

Here we also impose a bounded support on $[-1, 1]$: $K(u) = 0$ for $|u| \geq 1$. It follows that $K(u)$ has its maximum at $u = 0$. The kernel approximation of the score distribution is

$$\widehat{f}_h(P(X)) = \frac{1}{n} \sum_{i=1}^{n} K_h \left(P(X) - P(X_i)\right) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K \left(\frac{P(X) - P(X_i)}{h} \right),$$

where $K_h(\cdot) = h^{-1}K(\cdot/h)$. The number of neighbours included in the weighting is dependent on the bandwidth h, which determines the smoothing of the kernel and is a crucial parameter for the estimation. There are formulas for "optimal bandwidth" for e.g. symmetric distributions, but the easiest way is to make a simple sensitivity analysis with varying h. The kernel used here is the so called biweight (or quartic) kernel,

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

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

To sum up, our weight function will be

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

Heckman et al. (1998) show that the fundamental identification criterion of the matching method for estimating 1 is

$$E(Y_0|D=1,X) = (Y_0|D=0,X),$$

whenever both sides of the expression are well defined, and for both sides to be well defined it is necessary to condition on the common support S of both the treatment group and the non-treated:

$$S = \text{Supp}(X|D=1) \cap \text{Supp}(X|D=0).$$

This region of common support is estimated using the same kernel estimator as described above:

$$\widehat{f}_h(P(X_i)|D=0) = \frac{1}{N_0 \cdot h}\sum_{k=1}^{N_0} K\left(\frac{P(X_i)-P(X_k)}{h}\right), \quad \forall\, i \in I_1$$

The recommendation in Todd (1999) is to find the 1-2 % quantile of this distribution and discard the corresponding values of $P(X_i)$, i.e. where there are no close matches in the non-treated group. Of course, the amount of observations that should be discarded is dependent of the application, and might in this case seem to be small. We have used the higher number, 2 %.

The distributions of the productivity, the propensity scores (overall and for treated and non-treated respectively) are shown in figure 1, page 6. The

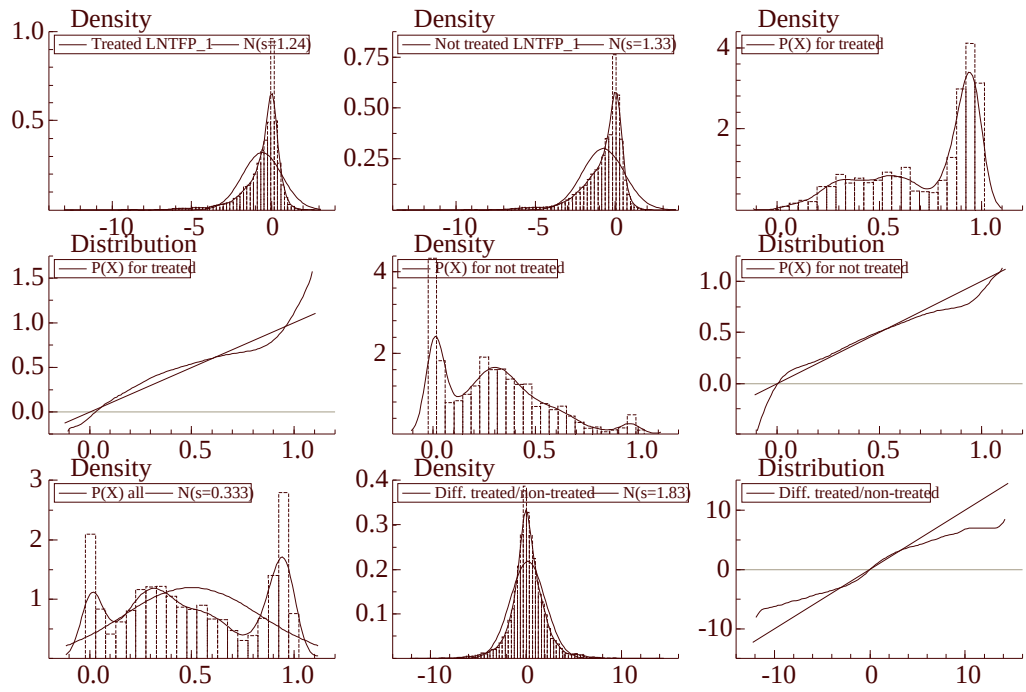

Figure 1: Densities and QQ-plots against the normal distribution, showing the productivity and propensity score for the treated (high accessibility) and non-treated (low accessilibity) in the service sector in Scania. The two bottom-most right figures concern the unweighted difference between the treated and non-treated groups.

productivity densities for treated and non-treated are quite similar, which is also shown in that the difference (bottom-most middle) centers around approximately zero. The propensity scores for treated and non-treated are markedly skewed to each end of the $[0, 1]$ interval, but both contain values for most of the spectrum except around zero (treated) and around 0.8–0.9 (not treated).

The common support of the service sector is presented in figure 2, page 7. Although there is a "hump" around zero, there are values evenly distributed along the axis up to the maximum value.

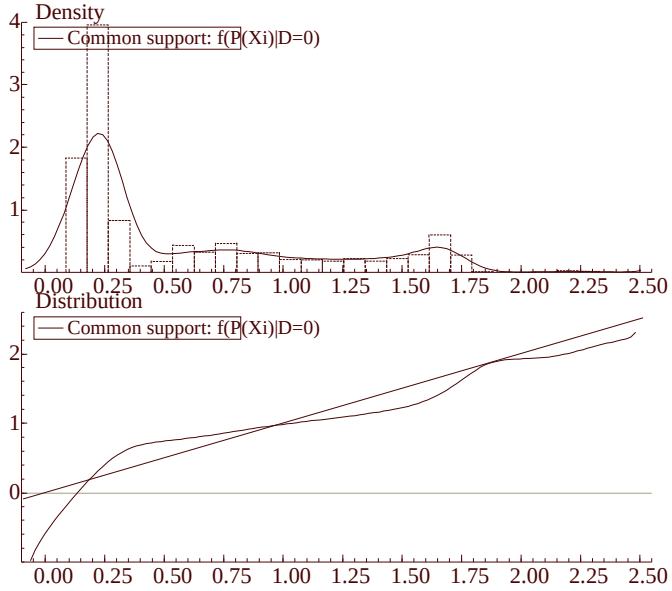

Figure 2: The common support of $f(P(X)|D = 1)$ and $f(P(X)|D = 0)$.

3.2 Weighted means

Another estimator for the average treatment effect is proposed by Horvitz & Thompson (1952) and Rosenbaum (1987), where the observations are weighted by the inverse of the probability of receiving the treatment actually received:

$$E\left(\frac{YD}{P(X)}\right) = E(Y_1), \quad E\left(\frac{Y(1-D)}{(1-P(X))}\right) = E(Y_0),$$

which can be used to estimate $E(Y_1 - Y_0)$.

4 Variables

4.1 Accessibility

Accessibility is a notion that has many different interpretations and could be expressed in many different ways (see e.g. Baradaran 2001, Handy & Niemeier 1997). In this paper we will use the expected utility in a random utility maximization framework as a measure of accessibility. The logsum variable is a measure of accessibility that is consistent with socio-economic

welfare theory (Williams 1977). It is a natural choice in the framework of the nested logit model, since it is the weighted expectation of the maximum utility an individual can derive from a set of choice alternatives, which is used on each level of the nest in order to summarize the utility in the levels below. In Baradaran (2001) it is called a "consumer surplus" approach. The definition is

$$A_i^n = \max_{i,j \in L} U_{j|i}^n,$$

where:

n is an individual,

j is the destination,

i is a given start node,

L is the total set of zones,

$U_{j|i}^n$ is the random utility of n, $U_{j|i}^n = v_j^n - c_{ij}^n + \varepsilon_{ij}$, where

v_j^n is the attractiveness of j to n,

c_{ij}^n is the cost of travel between i and j, and

ε_{ij} is the random, unobservable part of the utility

If ε_{ij} are independent and identically Gumbel distributed[1], the accessibility for individual n in zone i is:

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

where μ is a positive scale parameter – hence the term "logsum".

The accessibility variable used in this study is calculated by the SAMPERS system (the regional model for Scania and part of Denmark) (SIKA 2000). SAMPERS first calculates the travel demand pattern for the finest geographical level available, SAMS areas[2], then the accessibility measure is

[1]For the Gumbel distribution, see e.g. (Sydsæter, Strøm & Berck 1999).

[2]The SAMS zoal system is created by Sweden Statistics. SAMS stands for Small Area Market Statistics and comprise about 1000 residents per area. The zones are therefore smaller in the central cities and larger on the countryside. The approximate number of SAMS zones in Sweden is 9500; in Scania there are 1410 SAMS zones.

calculated on the finest zonal level, which are aggregates of a few SAMS zones. The number of zones in Scania is 948. The accessibility values are coded on to the firm dataset by the SAMS area code.

In SAMPERS, travel demand is calculated for only one trip purpose at a time. In this paper we use the work-trip purpose, since this is the closest proxy for the accessibility of firms to the labour market. In reality it measures the accessibility of workers to workplaces. The utility functions U_{ij}^n in the SAMPERS regional worktrips model include 35 variables (in total for all modes), e.g. number of employees as the main attraction variable.

4.2 Productivity

Productivity measures express the relationship between inputs and outputs in a production unit (be it a firm, municipality or a region). The concept of productivity incorporates both efficiency, i.e. proximity to the production frontier ("best practice" or most efficient firm) and scale effects, i.e. deviation from the optimal scale of the business in terms of output. The total factor productivity is defined as the ratio of output volume divided by the sum of the input volumes, weighted by their cost shares:

$$TFP = \frac{y}{\sum_i s_i x_i}, \tag{2}$$

where y is output volume, s_i is the cost share of input i, and x_i is input volume i ($s_i = w_i x_i / \sum_i w_i x_i = w_i x_i / C$). Taking logarithms, we get

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

which is the variable used in this work. The input factors considered here are labour, capital and material. The output volume and material volumes are calculated by dividing the monetary values of turnover and material cost by industry-specific indices during the time period of the study. For the output volumes, producer price indices (PPI) from Sweden Statisticsare used. For the material factor input, a mix of PPI, an implicit index based on GDP, and consumer price indices (CPI) were employed. The GDP-based indices and the CPI's were used for the service sectors, where PPI is missing.

These indices were weighted by an input-output matrix for Scania in order to get the right composition of the input for the different industries. This input-output matrix is the same as the one used in the Swedish regional

	N	Range	Minimum	Maximum	Mean		Std. Deviat
	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic
LNTFP_1	23294	16.62885	-13.33175	3.2971	-0.758496	0.008358	1.275676
PRE_3_1	22580	0.999982	1.19E-05	0.999994	0.510831	0.002105	0.316342
TILLG1_1	22148	3.071	8.829	11.9	10.51212	0.004154	0.618137
ANST_1	24873	3309.444	0	3309.444	7.91184	0.232612	36.68565
OMSATT_1	24873	7506942	-18783.28	7488159	12383.1	562.1497	88657.62
HIGHT_1	22148	1	0	1	0.506908	0.003359	0.499964
Valid N (listwi	15588						

Figure 3: Descriptive statistics of some firm specific variables, plus accessibility (TILLG_1) and propensity score (PRE_3_1). LNTFP_1 – log of productivity, ANST_1 – employees, OMSATT_1 – turnover, HIGHT_1 – treatment dummy.

economic model RAPS (Regional Analysis and Forecasting System), but it was aggregated to 33 branches according to available indices.

The index approach implies that in 1990, the number of input and output goods are the same as the monetary values, while they change in the following years according to the development in the different industries.

In the case of capital and labour, the inputs are simply the total debts and the number of full time worker equivalents during a year.

The cost share for capital is the sum of net financial cost and depreciation divided by total cost, the other cost shares simply the ratios of labour cost and material cost to total cost. Since the cost shares are values in the interval $[0, 1]$, it might seem odd to bring together variables with such different scale and dimension as total debts and number of workers in the same sum. However, a test with normalized variables did not show any remarkable difference of the distribution, only a shift along the axis.

Some descriptive statistics of some of the variables in the dataset is shown in table 3

4.3 Propensity score

The propensity score was obtained through the estimation of a binary logit model, and the use of the predicted probabilities for participation in group HIGH. The logit model was estimated using both location specific and individual specific variables. As mentioned, the choice alternatives were represented by "high" and "low" accessibility, where the boundary was the median of the accessibility variable. Not surprisingly, the location specific variables had the largest explanatory power – these were a dummy for the largest city Malmo, the number of workers in the area, built-up density, the number of

workplaces, a tourist area dummy, ratable value of real estate. The individual specific variables that had the highest explanatory power were, in order: branch affiliation, ownership, turnover, equity and net investments. The final rate of correct predictions of the logit model was 78 %. For a detailed description of the logit model, see the Appendix on page 18.

5 Results

5.1 Propensity score matching

The results of the propensity score matching (PSME) for 24 branch aggregates are presented in figure 4, page 12. As shown, the confidence limits are in every instance including the zero, so no statistical inference can be drawn upon the result. Not even in the branch categories with more than 1,000 observations in the treatment group (aggrID 11–12 and 19) there is any sign of a larger deviation of PSME from 0, or smaller standard deviation. The results on aggregate levels (total, manufacturing and service) are presented in figure 5, page 12. The density of the PSME for the total dataset is shown in figure 6, page 13.

A sensitivity analysis was performed on the total dataset with regard to the kernel bandwidth h. The result is presented in figure 7.

5.2 Weighted mean

The result of the weighted mean estimator is presented in table 8, page 14. In terms of efficiency, this estimator is far less efficient (standard deviations are much higher).

6 Discussion

We have not found any significant effect of a location in a high accessibility area on productivity of the firms in Scania, neither on an aggregate nor on a branch-divided level. This motivates a discussion about whether the assumptions of the method are fulfilled or not. There are several issues one could look deeper into: 1) Is there really independence in the "choice" between location in high or low accessibility areas? 2) Are not firms competing for space in this context, thus imposing restrictions on each others possibilities to choose area? 3) Can we disregard the fact that the conditioning variables X are actually data from the same period of study during which we study the

aggrID	obs	treated	PSME	std.dev.	CLO (2.5%)	CLH (2.5%)
0	850	160	0.025432	1.1683	[-2.4883	1.7788]
1	436	85	0.15725	1.3274	[-4.19	1.802]
2	280	110	0.050019	0.83745	[-1.9653	1.1598]
3	151	58	0.10831	0.93031	[-1.5234	1.7055]
4	557	335	-0.01385	1.0567	[-2.2803	1.5771]
5	387	134	0.49953	1.2704	[-2.0633	2.1792]
6	637	233	-0.044917	0.91568	[-1.7228	1.3358]
7	664	222	0.03838	0.98356	[-2.3493	1.3438]
8	379	223	0.4234	1.0651	[-2.0761	1.9318]
9	2350	849	-0.17127	1.1932	[-3.3143	1.307]
10	1001	386	0.15238	0.88118	[-2.2141	1.2037]
11	3338	1675	0.02146	0.96492	[-2.9765	1.0602]
12	3287	1514	-0.11762	0.88952	[-2.1892	0.86654]
13	1144	550	-0.22867	1.0428	[-2.6106	1.2731]
14	1212	400	0.042902	1.2201	[-2.3385	1.8107]
15	360	168	-0.10356	1.2013	[-2.6555	1.0921]
16	319	139	0.14278	2.3449	[-3.5938	4.2642]
17	309	174	-0.9766	1.9518	[-5.3807	1.6797]
18	875	448	0.31752	1.2199	[-2.7948	2.1336]
19	4212	2411	0.14516	1.2897	[-3.2581	1.7405]
20	209	112	0.14351	0.89998	[-1.5483	1.4868]
21	951	488	-0.22298	1.1966	[-3.1119	1.4971]
22	329	159	0.75222	1.2554	[-1.6106	2.8708]
23	385	180	-0.12283	1.1983	[-2.8323	1.4073]

Figure 4: Propensity score matching estimator for different branch aggregates (for the aggregation key, see the Appendix, page 20). The largest absolute effect is for branch 17 "Other real estate activities", although in the "wrong direction" (i.e. high accessibility is "bad" for this branch). Next largest effects, with positive sign, have 22 "Membership organizations, other service etc." and, rather unintuitively, 5 "Chemical manufacturing". CLO – lower confidence limit, CLH – higher confidence limit.

aggrID	obs	treated	PSME	std.dev.	CLO (2.5%)	CLH (2.5%)
all	24630	11217	0.022764	1.217881	[-3.35147	1.4679]
manuf	6691	2409	0.010063	1.1116	[-2.6236	1.4485]
service	17931	8804	0.018006	1.2459	[-3.6062	1.4609]

Figure 5: Result of the PSME on aggregate levels. "Manufacturing" includes agriculture, mining and construction as well.

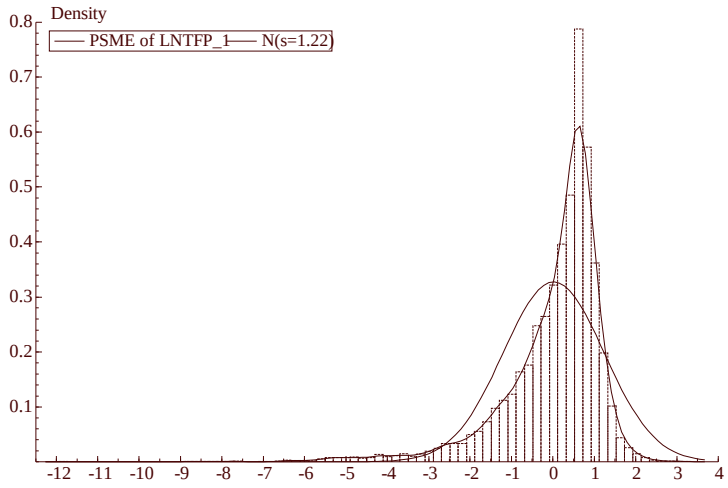

Figure 6: The propensity score matching estimator for the pooled dataset, a result typical for all branch subdivisions as well. Even if the median is on the positive side in this left-skewed distribution, the mean is very close to zero (0.023).

h	LNTFP_1	Std.dev.	CLO (2.5%)	CLH (2.5%)
0.05	0.024038	1.220032	-3.34387	1.48137
0.099877	0.022764	1.217881	-3.35147	1.4679
0.2	0.033605	1.214406	-3.30764	1.47038

Figure 7: Sensitivity analysis of PSME with regard to bandwidth h. The variations are small compared to the standard deviation and confidence limits, and do not alter the conclusions.

aggrID	treated	mean	std.dev.	CLO (2.5%)	CLH (2.5%)	CLO (5%)	CLH (5%)
0	160	-4.009	12.492	-20.408	4.3914	-16.2	3.094
1	85	-1.5313	5.7456	-17.174	3.6049	-12.2	3.224
2	110	-0.2927	2.8402	-7.5003	3.3736	-5.02	2.97
3	58	-0.4367	6.4853	-12.759	13.931	-6.66	3.853
4	220	-2.3477	12.916	-25.05	4.6455	-7.35	3.042
5	134	7.9595	40.174	-12.546	185.6	-10	25.78
6	233	-0.4311	2.8195	-6.5939	4.4447	-4.93	3.382
7	222	-1.2796	6.1398	-11.52	3.0887	-7.12	2.41
8	155	-6.0734	28.815	-61.045	3.7273	-32.1	3.221
9	849	-0.7509	7.7971	-13.503	6.3615	-8.9	4.839
10	386	-0.0693	4.4811	-7.3016	8.3448	-3.71	3.992
11	1654	-0.5863	12.293	-7.9786	9.3306	-4.65	5.229
12	1514	-0.1599	3.6327	-6.5447	4.0078	-3.32	3.154
13	550	1.9517	23.138	-6.4971	6.2155	-4.39	4.965
14	400	-2.1298	9.1267	-19.89	6.2572	-13.7	4.011
15	168	-0.6185	7.538	-12.12	10.383	-6.36	4.614
16	139	-0.5277	9.1952	-24.625	9.4066	-17.1	8.949
17	135	0.7837	5.4241	-6.717	12.195	-5.95	8.835
18	422	-1.172	8.2065	-11.104	8.1667	-7.52	5.287
19	1786	-0.8543	13.901	-9.5917	7.5449	-6	5.153
20	96	-2.238	15.08	-7.8983	3.5532	-6.7	3.018
21	462	0.3266	5.3557	-6.4065	9.1876	-4.71	6.569
22	159	0.1869	5.4127	-9.851	7.4955	-6.18	4.828
23	180	-0.1864	3.8908	-11.18	6.3117	-7.66	5.553

Figure 8: Results from the weighted mean estimator. All confidence intervals encompass zero, so there is no detectable effect of the treatment.

productivity, and consequently not strictly "pre-treatment variables" (and the productivity not strictly "post-treatment"? 4) A more methodological question, raised by Heckman et al. (1998), is: Can X be approximated by $P(X)$?

7 Conclusions

We have implemented the propensity score matching method on a number of firms, representing the whole spectrum of economic activity, in order to establish a connection between accessibility and productivity. The results here do not give any reason to believe that there are such links. Of course, some doubt can be cast upon the specification of the test. One possible refinement could be to separate observations before and after (i.e. taking non-treatment observations from the beginning of the panel period, and treatment observations from the end); another could be to investigate the change in productivity from year to year (Divisia or Tornquist index). Furthermore, in a recent paper Imbens (2000) proposes a "dose-response model" for causal effects, which would suggest to divide the accessibility variable in more groups than just two (implying a multinomial logit model for the propensity score).

References

Aschauer, D. A. (1989), 'Is public expenditure productive?', *Journal of Monetary Economics* **23**, 177–200.

Baltagi, B. H., Song, S. H. & Jung, B. C. (2001), 'The unbalanced nested error component regression model', *Journal of Econometrics* **101**, 357–381.

Baradaran, S. (2001), The Baltic Sea Region as a part of Europe. GIS Analyses of the Transport Infrastructure and Accessibility. Licentiate thesis, Dept. of Infrastructure and Planning, KTH, Royal Institute of Technology, Stockholm, TRITA-IP FR 01-86.

Chandra, A. & Thompson, E. (2000), 'Does Public Infrastructure Affect Economic Activity? Evidence from the Rural Interstate Highway System', *Regional Science and Urban Economics* **30**, 457–490.

Cox, D. R. & Snell, E. J. (1989), *The Analysis of Binary Data*, 2 edn, Chapman and Hall, London.

Handy, S. L. & Niemeier, D. A. (1997), 'Measuring accessibility: an exploration of issues and alternatives', *Environment and Planning A* **29**, 1175–1194.

Härdle, W. & Linton, O. (1994), Applied nonparametric methods, *in* R. F. Engle & D. L. McFadden, eds, 'Handbook of Econometrics', Vol. 4, Elsevier Science, chapter 38, pp. 2295–2339.

Heckman, J. J., Ichimura, H. & Todd, P. (1998), 'Matching as an econometric evaluation estimator', *Review of Economic Studies* **65**, 261–294.

Holtz-Eakin, D. (1994), 'Public-sector capital and the productivity puzzle', *Review of Economics and Statistics* **76**, 12–21.

Horvitz, D. & Thompson, D. (1952), 'A generalization of sampling without replacement from a finite population', *Journal of The American Statistical Association* **47**, 663–85.

Imbens, G. W. (2000), 'The role of the propensity score in estimating dose-response functions', *Biometrika* **87**(3), 706–710.

Moreno, R., Artis, M., López-Bazo, E. & Suriñach, J. (1997), 'Evidence on the complex link between infrastructures and regional development', *International Journal of Development Planning Literature* **12**, 81–108.

Munnell, A. H. (1990), 'Why has productivity growth declined? Productivity and public investment', *New England Economic Review* pp. 3–22.

Nagelkerke, N. J. D. (1991), 'A note on a general definition of the coefficient of determination', *Biometrika* **78**, 691–692.

Rosenbaum, P. (1987), 'Model-based direct adjustment', *Journal of The American Statistical Association* **82**, 387–94.

Rosenbaum, P. R. & Rubin, D. B. (1983), 'The central role of the propensity score in observational studies for causal effects', *Biometrika* **70**(1), 41–55.

Rosenbaum, P. R. & Rubin, D. B. (1985), 'Constructing a control group using multivariate matched samplingmethods that incorporate the propensity score', *The American Statistician* **39**(1), 33–38.

SIKA (2000), 'Sampers – Användarmanual (User's Manual). The Swedish National Model System 2000', Stockholm, Sweden.

Sydsæter, K., Strøm, A. & Berck, P. (1999), *Economists' Mathematical Manual*, 3 edn, Springer, Berlin.

Todd, P. (1999), A practical guide to implementing matching estimators. http://athena.sas.upenn.edu/∼petra/prac.pdf.

Williams, H. C. W. L. (1977), 'On the formation of travel demand models and economic evaluation measures of user benefit', *Environment and Planning A* **9**, 285–344.

8 Appendix

8.1 The binary logit model

Here the details of the logit model for the propensity score are presented, see tables in figures 9 and 10.

		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 9: The binary logit model for propensity score estimation. The variables are, in order (L means location specific): ownership, turnover, equity, net investments, branch affiliation, number of workplaces, built-up density, number of workers, large university, small university, regional hospital, county hospital, larger shopping centre, superstore, hyperstore, tourist area dummy, touristic point dummy, real estate ratable value, other ratable value, chief town in the municipality, chief town in the county (i.e. Malmö).

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

Figure 10: Model summary of the logit model. The R square measures (Cox & Snell 1989, Nagelkerke 1991) are goodness-of-fit measures in the interval [0,1] and should be as high as possible, while the –2·log-likelihood should be minimized. The overall percentage correctly predicted individuals was actually somewhat higher in step 16, before some individual-specific variables were entered in the model. HIGHT_1 – treatment group.

8.2 Aggregation key

Below is the key between aggregation ID in the tables in 4. SNI stands for the Swedish standard for branch subdivisions, with international equivalents like NACE (European Union) and ISIC (United Nations).

****

Aggregation id: 0

Selected SNI group(s):

1 Agriculture, hunting and related service activities

****

Aggregation id: 1

Selected SNI group(s):

2 Forestry, logging and related service activities

5 Fishing, operation of fish hatcheries and fish farms; service activities incidental to fishing

10 Mining of coal and lignite; extraction of peat

11 Extraction of crude petroleum and natural gas; service activities incidental to oil and gas extraction, excluding surveying

12 Mining of uranium and thorium ores

13 Mining of metal ores

14 Other mining and quarrying

40 Sawmilling and planing of wood

41 Manufacture of products of wood, cork, straw and plaiting materials

201 Manufacture of pulp

209 Manufacture of paper and paper products

211 Electricity, gas, steam and hot water supply

219 Collection, purification and distribution of water

****

Aggregation id: 2

Selected SNI group(s):

15 Manufacture of food products and beverages

16 Manufacture of tobacco products

****

Aggregation id: 3

Selected SNI group(s):

17 Manufacture of textiles

18 Manufacture of wearing apparel; dressing and dyeing of fur

19 Tanning and dressing of leather; manufacture of luggage, handbags, saddlery, harness and footwear

****

Aggregation id: 4
Selected SNI group(s):
22 Publishing, printing and reproduction of recorded media
****
Aggregation id: 5
Selected SNI group(s):
23 Manufacture of coke, refined petroleum products and nuclear fuel
24 Manufacture of chemicals and chemical products
25 Manufacture of rubber and plastic products
26 Manufacture of other non-metallic mineral products
****
Aggregation id: 6
Selected SNI group(s):
27 Manufacture of basic metals
28 Manufacture of fabricated metal products, except machinery and equipment
****
Aggregation id: 7
Selected SNI group(s):
29 Manufacture of machinery and equipment n.e.c.
34 Manufacture of motor vehicles, trailers and semi-trailers
35 Manufacture of other transport equipment
36 Manufacture of furniture; manufacturing n.e.c.
37 Recycling
****
Aggregation id: 9
Selected SNI group(s):
45 Construction
****
Aggregation id: 10
Selected SNI group(s):
50 Sale, maintenance and repair of motor vehicles and motorcycles; retail sale of automotive fuel
****
Aggregation id: 11
Selected SNI group(s):
51 Wholesale trade and commission trade, except of motor vehicles and motorcycles
****
Aggregation id: 12
Selected SNI group(s):

52 Retail trade, except of motor vehicles and motorcycles; repair of personal and household goods

****

Aggregation id: 13

Selected SNI group(s):

55 Hotels and restaurants

****

Aggregation id: 14

Selected SNI group(s):

60 Land transport; transport via pipelines

61 Water transport

62 Air transport

****

Aggregation id: 15

Selected SNI group(s):

63 Supporting and auxiliary transport activities; activities of travel agencies

64 Post and telecommunications

****

Aggregation id: 16

Selected SNI group(s):

702 Real estate activities with own property

****

Aggregation id: 17

Selected SNI group(s):

709 Other real estate activities

****

Aggregation id: 18

Selected SNI group(s):

71 Renting of machinery and equipment without operator and of personal and household goods

72 Computer and related activities

73 Research and development

****

Aggregation id: 19

Selected SNI group(s):

74 Other business activities

****

Aggregation id: 20

Selected SNI group(s):

80 Education

****
Aggregation id: 21
Selected SNI group(s):
85 Health and social work
****
Aggregation id: 22
Selected SNI group(s):
90 Sewage and refuse disposal, sanitation and similar activities
91 Activities of membership organizations n.e.c.
93 Other service activities
****
Aggregation id: 23
Selected SNI group(s):
92 Recreational, cultural and sporting activities