ECON 881 Paper

Zafer Rabin Saba

Investigating The Change in Inventor Productivity After Moving: Evidence from US Patent Data

1. Introduction

Inventors make up a unique class of high skilled workers. They provide important benefits to society and their work is an important driver of human progress. The availability of patent data makes it possible to take a close look at their productivity. Analysis of inventor productivity can also help understand the determinants of productivity for the overall skilled worker class, which is usually hard to measure.

This paper focuses on the causal effect of moving on the productivity of inventors, using US patent data with approximately 1.3 million different patents produced by 587,276 different inventors between 2004 and 2020. Using German data, Hoisl (2007) shows that moving to a different firm has a positive effect on the productivity of inventors on average but does not consider moves to a new location. This paper investigates moves to a different state together with moves to a different firm to understand the separate effect of moving to a new location, and the interaction between these. I confirm Hoisl (2007) also holds with US Patent data, i.e., changing firms leads to an increase in productivity. I find no evidence that moving to a new location leads to an increase in productivity after controlling for firm change. I also find that the estimated coefficient for the interaction between firm change and location change is negative.

2. Data

A) Raw Data

Data from the publicly available Patentsview database of US patent data was used for this paper. Specifically, 7 different datasets from this database were used:

1) **Patent:** Contains characteristics of granted patents
2) **Inventor:** Contains disambiguated data about characteristics of inventors
3) **Location:** Contains disambiguated location data
4) **Patent_Inventor:** Connects Patent, Inventor, and Location tables
5) **Uspatentcitation:** Contains all the citations of a patent to other patents
6) **Assignee:** Contains information about the owners of the patents. These are companies, universities, research labs, or other organizations that employ the inventor.
7) **Assignee_Patent:** Connects Assignee and Patent tables.

I used these datasets to analyze the effect of moving on the productivity of inventors between 2006 and 2015. I have three main reasons to choose this period: Firstly, I want to make sure that the period is large enough to have a good enough sample size. Secondly, I do not want the dates to be outdated. Thirdly, I want to allow 5 years for each published patent to get citations so that their value can be accurately measured, which explains why I stop at 2015. While the effect of moving is estimated for these periods, inventors are tracked between 2004 and 2020 to have productivity information for all of them in the 2 years before the move period and the 5 years after the move period.

B) Data Processing

The above datasets were merged to get a final dataset that contains all the necessary information on the inventors, which are going to be explained later in the Variable Description part.

The state information in each patent was used to track the location of inventors. If an inventor published a patent in a different state compared to his/her previous patent, then this was interpreted as a location move.

Similarly, the assignee data for the patents were used to track the firms that inventors work for. Patents that belonged to organizations other than firms were removed from the data. If an inventor published a patent that was assigned to a different firm compared to his/her previous patent, this was coded as a change of firm for the inventor.

Unfortunately, it is not possible to obtain the exact move date from the patent data, therefore, the move date was estimated as the average of the dates between the last patent published before moving, and the first patent published after moving.

Lastly, citation data was used to estimate the productivity of inventors. According to Harhoff et al. (2003), the number of citations is positively related to the value of a patent. Therefore, to measure the productivity of inventors in a period, I looked at the patents they published during that period and calculated the total citations these patents get after 5 years of publication.

There were also some edge cases in the data. Some of the patents were invented by a single inventor but belonged to multiple firms, possibly because of collaboration between firms. Such patents were removed from the data as it is not clear whether they indicate a firm move

or not for the inventors. Additionally, some inventors published two different patents in different locations or firms in the same day, and such a case was also not interpreted as a move.

C) Variable Descriptions

My dependent variable of interest y is the productivity of inventors in the 5-year period after moving, which is the average yearly citations they get. I am interested in the effect of 3 different independent variables, location_mover, which is equal to 1 for inventors who change their state at any point between 2006 and 2015, firm_mover, which is equal to 1 for inventors who change their firm at least once between 2006 and 2015, and location_moverXfirm_mover, which is the interaction of these variables. My other independent variables are male (equal to 1 for males), productivity_before which is the average yearly citations in the 2 years before moving, and inventor experience, which is the time passed after the inventor published their first patent. Unfortunately, age or birth year data were not available, but inventor experience can be used as a similar measure.

3. Methodology

A) Design

Consider an arbitrary inventor with n distinct patents. The set of patents for this arbitrary inventor can be denoted by $p = \{p_1, p_2 \dots p_n\}$, the corresponding locations (The US state that the patent was published in) of these patents can be denoted by $l = \{l_1, l_2 \dots l_n\}$ and the firms that are assigned these patents can be denoted by $f = \{f_1, f_2 \dots f_n\}$. The set $a = \{a_1, a_2 \dots a_n\}$ shows whether a new patent indicates a location move or not. a_1 will always be

0, and a_i is equal to 1 if $l_i \neq l_{i-1}$, otherwise, it is equal to 0 for $i \in \{2 \ldots n\}$. The set $b = \{b_1, b_2 \ldots b_n\}$ shows whether a new patent indicates a firm move or not, and is calculated very similar to the location move. b_1 will always be 0, and b_i will equal 1 if $f_i \neq f_{i-1}$, otherwise, it is equal to 0 for $i \in \{2 \ldots n\}$. The set $c = \{c_1, c_2 \ldots c_n\}$ shows whether a new patent indicates any move at all. c_i is calculated as max (l_i, f_i) for $i \in \{1 \ldots n\}$. Note that by definition, $c_1 = 0$ for all inventors. If $a_i = 0$ for all $i \in \{1 \ldots n\}$, this inventor is classified as a location stayer, and if $a_i = 1$ for any $i \in \{1 \ldots n\}$, this inventor is classified as a location mover. Similarly, if $b_i = 0$ for all $i \in \{1 \ldots n\}$, this inventor is classified as a firm stayer, and if $b_i = 1$ for any $i \in \{1 \ldots n\}$, this inventor is classified as a firm mover. If an inventor changes both firm and location at the same time, i.e., $a_i = 1$ and $b_i = 1$ for some $i \in \{1 \ldots n\}$, this inventor is classified as both a firm and location mover. However, I only consider the first move of an inventor to classify them. For instance, if the first move of an inventor is to change only his/her firm in time j and this inventor only changes location in time k, where j < k, then this inventor will be classified as a firm mover but not a location mover as the firm move happened first. This is because my analysis requires a unique move date for each inventor, and I believe the first move to be the most important determiner of productivity, although further research is needed to confirm this empirically. Lastly, if $c_i = 0$ for all $i \in \{1 \ldots n\}$, this inventor will be in the control group and called a mover, and if $c_i = 1$ for any $i \in \{1 \ldots n\}$, this inventor will be in the treatment group and called a stayer.

My overall research design is similar to Hoisl (2009) where there are pre-move and post-move periods, and the effect of moving was estimated after matching between the different groups of inventors. My pre-move period is 2 years in length whereas the post-move period is 5

years. These were chosen as such to ensure that I get the recent productivity of inventors before moving, while also looking at the long-term effect of moving as there might be an adjustment period for inventors. Note that I do not include the productivity of an inventor in the move year in any of the pre-move or post-move periods as inventors would spend some of this year under their previous location or employer but would spend the other portion in their new location or firm.

B) Different Specifications and Matching

I use two different merged datasets for this paper. In the first one, all inventors who publish a patent are observed. This dataset faces some issues while matching, which are explained later. Then, I suggest an alternative dataset that is the same as the first one, with only a minor change that inventors without any citations in the previous 2 years before the move date are removed so only inventors whose patents received at least 1 citation are observed. I argue that the second dataset gives more reliable results.

For each dataset, I present a naïve OLS regression that is done without any matching between movers and stayers. Then, I do causal modeling by using different matching methods between movers and stayers. In the absence of matching, the OLS regression suffers from endogeneity issue because there might be reverse causality (differences in productivity might be the reason for inventors moving), or there might be other variables that affect who moves, which in turn affect productivity.

Movers were matched with stayers to ensure movers and stayers have similar observable characteristics in the pre-move period, and their only difference was the treatment, which is moving to a new location or a new firm. Coarsened exact matching and propensity

score matching were both tried for matching inventors who moved with the ones who did not. The matching was done based on the productivity in the previous two years, inventor experience, and the gender of the inventor. All these values were observed in the same year for a pair of matched inventors. For example, an inventor who moved in 2018 was matched with an inventor who did not move, but both had similar values in 2018 for average productivity in the previous two years, and experience, while also having the same gender.

Propensity score matching did not converge in the first dataset for both logit and probit models. Specifically, the first stage of estimating propensity scores that estimate the probability of moving based on observed covariates worked normally, however, matching participants based on their propensity score did not converge using the nearest neighbor matching algorithm. I suspect this is because of outliers that have 0 productivity before the moving period (i.e., they received 0 citations in the last 2 years) because coarsened exact matching also had issues with balancing the productivity in the pre-move period between movers and stayers. (More information about this on the next chapter). Because probit and logit models both did not converge, coarsened exact matching was used instead for the first dataset, but it did not result in a balanced match. Therefore, this dataset was not very useful for causal inference, but provides descriptive information.

Next, I do a similar analysis with the second dataset. I start with a naïve OLS, and then implement matching using CEM and logit propensity score matching. The non-convergence issue in the first dataset does not exist here, and a balanced dataset was obtained using propensity score matching. Therefore, I use this as my main specification.

C) Balance Statistics

Out of 587,276 inventors, 44,706 of them (7.6%) changed either their firm or location at least once. A key assumption of causal inference by matching is that the treatment and control groups have similar characteristics. In absence of that, we would run into similar endogeneity issues that are present in the naïve OLS model. Therefore, I look more deeply into this assumption by checking the balance of the dataset before and after matching.

i. First Dataset

The balance statistics between treatment and control groups for the first dataset can be seen below:

```
Call:
matchit(formula = treatment ~ year + productivity_last2_years +
    experience + male, data = i0, method = "cem", estimand = "ATT")

Summary of Balance for All Data:
                         Means Treated Means Control Std. Mean Diff. Var. Ratio eCDF Mean eCDF Max
year                        2010.2856     2010.9203         -0.2089     1.1259    0.0635   0.0947
productivity_last2_years       9.8365        3.6395          0.1403     4.3536    0.0112   0.2905
experience                     6.9006        5.1479          0.2116     1.1085    0.0439   0.2331
male                           0.9035        0.8704          0.1122          .    0.0331   0.0331


Summary of Balance for Matched Data:
                         Means Treated Means Control Std. Mean Diff. Var. Ratio eCDF Mean eCDF Max Std. Pair Dist.
year                        2010.2873     2010.2873          0.0000     1.0000    0.0000   0.0000          0.0000
productivity_last2_years       9.0639        4.5007          0.1033     1.6019    0.0087   0.2288          0.1144
experience                     6.8922        6.7790          0.0137     0.9797    0.0029   0.1108          0.0564
male                           0.9035        0.9035         -0.0000          .    0.0000   0.0000          0.0000

Percent Balance Improvement:
                         Std. Mean Diff. Var. Ratio eCDF Mean eCDF Max
year                               100.0        100     100.0    100.0
productivity_last2_years            26.4         68      22.3     21.2
experience                          93.5         80      93.3     52.5
male                               100.0          .     100.0    100.0

Sample Sizes:
              Control Treated
All           542570.    44706
Matched (ESS) 428430.9   44657
Matched       542097.    44657
Unmatched        473.       49
Discarded          0.        0
```

Figure 1, Balance Statistics for CEM with the First Dataset

Before matching, productivity in the last 2 years is the biggest difference between treatment and control groups. However, a match was found for almost all inventors with CEM

and a more balanced dataset was created as there were improvements in the percent balance for all the variables. However, there is still an important difference between the average productivity in the last 2 years between the treatment (9.06) and the control group (4.5) in the matched data. The difference is statistically significant using a 2-sided t-test, so a causal inference is likely to be unhealthy in this case.

ii. Second Dataset

Now, I move on to matching using the second dataset. I first use CEM, the results can be seen below:

```
Summary of Balance for All Data:
                         Means Treated Means Control Std. Mean Diff. Var. Ratio eCDF Mean eCDF Max
year                        2010.0617     2010.9606         -0.2899     1.1495    0.0899   0.1319
productivity_last2_years       15.6451       10.7599          0.0890     2.4131    0.0086   0.0935
experience                      8.2762        7.6952          0.0712     1.0616    0.0168   0.0448
male                            0.9100        0.8879          0.0771          .    0.0221   0.0221


Summary of Balance for Matched Data:
                         Means Treated Means Control Std. Mean Diff. Var. Ratio eCDF Mean eCDF Max Std. Pair Dist.
year                         2010.063     2010.0635          0.0000     1.0000    0.0000   0.0000          0.0000
productivity_last2_years       14.481       10.4840          0.0728     1.3228    0.0077   0.0965          0.2442
experience                      8.268        8.3117         -0.0054     1.0097    0.0013   0.0444          0.0602
male                            0.910        0.9100          0.0000          .    0.0000   0.0000          0.0000

Percent Balance Improvement:
                         Std. Mean Diff. Var. Ratio eCDF Mean eCDF Max
year                               100.0      100.0     100.0    100.0
productivity_last2_years            18.2       68.2      10.6     -3.2
experience                          92.5       83.9      92.2      0.9
male                               100.0          .     100.0    100.0

Sample Sizes:
                Control Treated
All            183520.    28108
Matched (ESS) 147922.2    28064
Matched       183143.    28064
Unmatched        377.       44
Discarded          0.        0
```

Figure 2, Balance statistics for CEM with the Second Dataset

Similar to CEM with the first dataset, the means for productivity in the last 2 years after matching are not close enough to claim a balanced dataset. Therefore, I try propensity score matching with logit instead:

```
Summary of Balance for All Data:
                         Means Treated Means Control Std. Mean Diff. Var. Ratio eCDF Mean eCDF Max
distance                        0.1454        0.1309          0.3134     1.4175    0.0967   0.1434
productivity_last2_years       15.6451       10.7599          0.0890     2.4131    0.0086   0.0935
experience                      8.2762        7.6952          0.0712     1.0616    0.0168   0.0448
male                            0.9100        0.8879          0.0771          .    0.0221   0.0221
year                         2010.0617     2010.9606         -0.2899     1.1495    0.0899   0.1319


Summary of Balance for Matched Data:
                         Means Treated Means Control Std. Mean Diff. Var. Ratio eCDF Mean eCDF Max Std. Pair Dist.
distance                        0.1454        0.1453          0.0004     1.0037    0.0000   0.0006          0.0005
productivity_last2_years       15.6451       14.9830          0.0121     1.1599    0.0032   0.0263          0.1090
experience                      8.2762        8.3179         -0.0051     0.9750    0.0015   0.0048          0.1439
male                            0.9100        0.9170         -0.0244          .    0.0070   0.0070          0.0728
year                         2010.0617     2010.0628         -0.0003     1.0151    0.0026   0.0083          0.0563

Percent Balance Improvement:
                         Std. Mean Diff. Var. Ratio eCDF Mean eCDF Max
distance                            99.9       98.9     100.0     99.6
productivity_last2_years            86.4       83.2      62.6     71.9
experience                          92.8       57.7      90.9     89.3
male                                68.4          .      68.4     68.4
year                                99.9       89.2      97.1     93.7

Sample Sizes:
          Control Treated
All        183520   28108
Matched     28108   28108
Unmatched  155412       0
Discarded       0       0
```

Figure 3, Balance Statistics with Logit, Second Dataset

This gives a much better match for productivity in the last 2 years, as the means are close to each other for treatment and control groups, and the Standard Mean Difference is lower than 0.05 threshold. This was a major problem in the other specifications. All the treated inventors (the ones who moved) were matched to a control one. QQ Plot for this can be seen below:

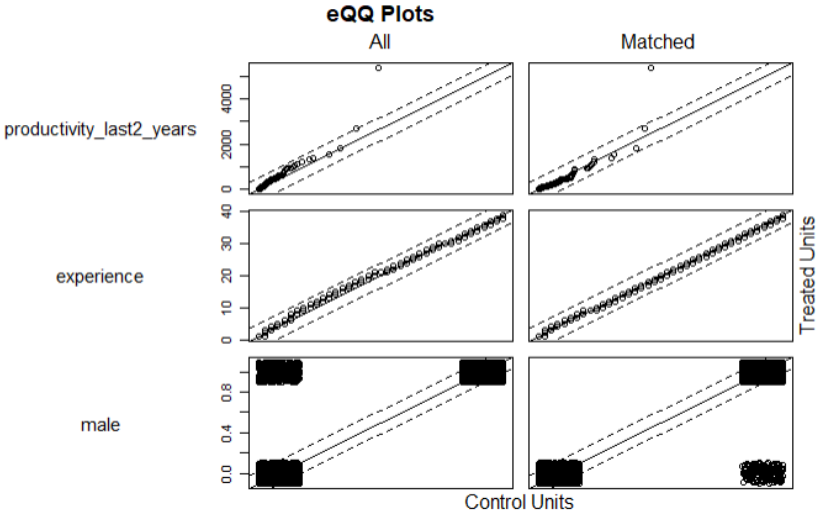

Figure 4, QQ plot for logit, second dataset

Overall, the control and treated units are scattered around the 45-degree line in the matched case for productivity in the last 2 years and experience, which indicates a balanced match. For gender, control units and treated units are again scattered close to each other. They are in the bottom right corner as expected because a very large percentage (89%) of all inventors are male. Additionally, using a two-sided t-test, the difference between the treatment and control group is not statistically significant for any of the variables. Considering all of these, I decide to use the second dataset with logit propensity score matching as my main model. However, I report the results for the other specifications as well in Chapter 5 because they provide information about different samples compared to the main model.

The results of first stage logistic regression for the main model can be seen below:

	Dependent variable:
	treatment
productivity_last2_years	0.002***
	(0.0002)
experience	0.008***
	(0.001)
male	0.221***
	(0.022)
Constant	-2.167***
	(0.022)
Observations	211,628
Note:	*p<0.1; **p<0.05; ***p<0.01
	Logit Propensity Score, First Stage

Table 1, Logit First Stage

Based on the first stage logistic regression in Figure 6, higher productivity in the pre-move period, more experience as an inventor, and being a male all increase the probability of an inventor moving to a new location or firm compared to their counterparts. All of these are statistically significant.

4. Model

Except for naïve OLS, the datasets used in the regressions only include matched inventors. After the final dataset is established, the following equation was estimated in all the specifications:

$$\begin{aligned}\text{productivity}_{\text{i}} = {} & \beta_0 + \beta_1 * location\_mover_{\text{i}} + \beta_2 * \text{firm}\_mover_{\text{i}} \\ & + \beta_3 * location\_mover * \text{firm}\_mover_{\text{i}} + \beta_4 * \text{experience}_{\text{i}} + \beta_5 * male_{\text{i}} \\ & + \beta_6 * \text{productivity\_before}_{\text{i}} + \text{e}_{\text{i}}\end{aligned}$$

where, e_{i} is the error term for inventor I, and the other variables were explained in part 2.C.

5. Results

A) First Dataset

	Dependent variable:	
	productivity_next5_years	
	(1)	(2)
productivity_last2_years	0.611***	0.486***
	(0.002)	(0.002)
experience	-0.062***	-0.029***
	(0.005)	(0.005)
male	0.529***	0.690***
	(0.121)	(0.128)
location_mover	1.608***	2.060***
	(0.443)	(0.416)
firm_mover	6.099***	6.121***
	(0.182)	(0.171)
location_moverXfirm_mover	-1.300**	-1.181**
	(0.540)	(0.506)
Constant	0.404***	0.447***
	(0.113)	(0.122)
Note:	*p<0.1; **p<0.05; ***p<0.01	
	First dataset. (1): Naive OLS. (2): OLS with CEM	

Table 2, Results for First Dataset

Overall, the results of Naïve OLS (1) and the OLS with CEM (2) are very similar to each other, which supports the issue that matching did not do enough to balance the first dataset. Therefore, these results should be interpreted as descriptive instead of causal. All the variates in the regression gave statistically significant results. Both regressions show that inventors who change their firm get approximately 6 more yearly citations on average, which is a much higher increase compared to inventors who change their state. In the second column, the interaction

term is barely statistically significant using the 0.05 p-value, and this significance goes away after using robust standard errors while other variables are still statistically significant with robust standard errors. The negative sign on the interaction term indicates that the combined effect is less than the sum of the two effects. Another interesting thing is the negative estimated coefficient for experience. A possible reason for this is that a linear model does not capture the full effect of experience. As inventors retire, their productivity becomes 0, but their experience continues to increase because of the way I defined it, which might be overshadowing the increase in productivity that comes from inventors spending more years in the workforce.

B) Second Dataset

	Dependent variable:		
	productivity_next5_years		
	(1)	(2)	(3)
productivity_last2_years	0.622***	0.473***	0.730***
	(0.002)	(0.003)	(0.005)
experience	-0.039***	-0.010	-0.031
	(0.012)	(0.010)	(0.032)
male	1.254***	1.150***	2.050**
	(0.307)	(0.294)	(0.941)
location_mover	1.228	1.384*	1.838
	(0.808)	(0.711)	(1.171)
firm_mover	7.648***	7.493***	8.544***
	(0.332)	(0.292)	(0.573)
location_moverXfirm_mover	-2.384**	-2.015**	-2.601*
	(0.985)	(0.866)	(1.418)
Constant	-2.155***	-0.616**	-5.675***
	(0.297)	(0.287)	(0.946)
Note:			*p<0.1; **p<0.05; ***p<0.01
	Using second dataset,(1): Naive OLS. (2): OLS with CEM, (3): OLS with Logit Propensity Score		

Table 3, Results for Second Dataset

The results of the regression with the second dataset are a little different. Here, although the estimated coefficient for location mover is positive for all 3 versions, it is not statistically significant in Naïve OLS and OLS with Logit. Additionally, the estimated coefficient for location mover is less than the first dataset in both. This shows that non-productive inventors are affected more positively by changing their location as the second dataset removed inventors with 0 citations in the previous 2 years.

The 3rd column is the main specification where propensity score matching was implemented in the first stage. While the effect of changing location is insignificant, changing firms is significant in the 0.01 level and leads to 8.5 more yearly citations on average. Meanwhile, the effect of experience is practically gone and is statistically insignificant. In addition, it is interesting that the interaction term is statistically significant in the 0.05 level while location_mover is not. Lastly, controlling for all the other factors, males get 2 more yearly citations on average in the post-move period.

6. Caveats

There are some caveats with the data and methodology that lead to biased or ungeneralizable results.

Most importantly, there might be unobserved variables that differ among matched movers and stayers, such as education and IQ. According to Akcigit, Pearce and Prato (2020), education is an important determiner of who innovates or not. Using Danish micro-data, they find that people with a PhD degree are approximately 20 times more likely to become inventors, and people with higher IQ are also more likely to be inventors after controlling for education. Further research is needed to understand the relationship of moving with education

and IQ, which would give an idea of the direction of bias because of the omission of these variables.

Secondly, because inventors are tracked by the patents they publish, inventors who did not publish anything after moving to a new firm or location are not observed. Such inventors are missing from the analysis. Therefore, this leads to an overestimation of the effect of moving on productivity. However, this is the only way to track inventors unless we match the patent data with something more comprehensive such as Social Security records.

Thirdly, there might be heterogeneity in the effect of moving based on experience, field, or productivity before moving. For instance, top inventors may benefit more from moving compared to lower-performing inventors, although movers and stayers were matched based on these characteristics.

Lastly, the results with matching are only representative of inventors who have close matches as other inventors are removed from the dataset. The Naïve OLS includes all inventors, but it cannot be used to make a causal argument.

7. Conclusion

To sum up, the location and firm changes of inventors and their productivity were tracked using patent data. Using propensity score matching with a logit model as my first stage, and then performing OLS regression on the matched dataset, I find that moving to a new firm increases the productivity of an inventor, confirming the previous research of Hoisl (2007) for the US case. I find no evidence that moving to a new state increases productivity, although there is a positive correlation between changing location and productivity in all the specifications.

For further research, it would be interesting to examine the differences in the effect of moving for inventors who move to big firms and inventors who move to small firms. Additionally, similar methodology could be used to investigate the effect of moving on the productivity of academic researchers. Like using patents to track inventors, academic researchers could be tracked by their publications.

References

Akcigit, U., Pearce, J., & Prato, M. (2021). Tapping into talent: Coupling Education and innovation policies for economic growth. https://doi.org/10.3386/w27862

Harhoff, D., Scherer, F. M., & Vopel, K. (2003). Citations, family size, opposition and the value of patent rights. *Research Policy*, *32*(8), 1343–1363. https://doi.org/10.1016/s0048-7333(02)00124-5

Hoisl, K. (2007). Tracing Mobile inventors—the causality between inventor mobility and inventor productivity. *Research Policy*, *36*(5), 619–636. https://doi.org/10.1016/j.respol.2007.01.009

Hoisl, K. (2009). Does mobility increase the productivity of inventors? *The Journal of Technology Transfer*, *34*(2), 212–225. https://doi.org/10.1007/s10961-007-9068-5