README

This README – file relates to the paper Aghion, Akcigit, Hyytinen and Toivanen "A year older, a year wiser (and farther from frontier): Invention rents and human capital depreciation".

This file consists of the following parts:

1. Matching the public inventor data with individual identifiers to enable matching with Finnish registry data
2. Command file structure
3. Matching the individual characteristics data from Statistics Finland with the inventor data.
4. Forming the matched estimation sample
5. Descriptive statistics analysis
6. Regression (DID) analysis.

We next detail each of these in turn. When it comes to variables, we concentrate on those variables used in i) matching the different data sets, ii) matching the treated individuals with controls, and iii) regression analysis. We write variable and file names using *italics*.

1. Matching the public inventor data with individual identifiers to enable matching with Finnish registry data

The information on inventors is based on OECD PATSTAT 2015 data. These are documented in *OECD REGPAT Database - July 2014.pdf* and *OECD_DSTI_DOC_2013_3.pdf*. The OECD data includes information on the inventors in each patent, their addresses, their employer, and the employer's address. All those are used in the matching process which is described in Table 1. Column 1 identifies the iteration round; column 2 the matching rule; column 3 the information in PATSTAT that is utilized; and column 4 the information in the Statistics Finland registry data that is used.

All in all, we match 89.6% of the inventors. 92% of matches are achieved in iteration round A, i.e., using rules #1-#11. In total there are four iteration rounds as follows (see first column in Table 1):

Iteration round A: all, 1978-2013 (data problems related on firm id's 1978-1987)

Iteration round B: Different employment periods available only for the period 2001-2012

Iteration round C: all, 1978-2012

Iteration round D: only done for years 1981-2012

This matching procedure produces the data set *patstat_19782013.dta* which contains the PATSTAT information and the Statistics Finland pseudonymized individual identifier *shnro*. The observation unit is an individual – patent combination.

Table 1. Matching PATSTAT data to Statistics Finland data

Iteration round	**Rule**	**PATSTAT**	**Population registry files**
A	**1**	Last name	Last name
A		First name (only the first)	First name

A		Firm id found from registry	Firm id (registry, last week of the year)
A		Postcode	Postcode
A		Street	Street
A		Street number	Street number
A		Application year	Registry year
A	**2**	Diff to rule 1: firm id from patent data base	
A	**3**	Diff to rule 1: no street number	
A	**4**	Diff to rule 1: firm id from patent data base	
A		Diff to rule 1: no street number	
A	**5**	Diff to rule 1: no firm id at all	
A	**6**	Diff to rule 1: no firm id at all	
A		Diff to rule 1: no street number	
A	**7**	Diff to rule 1: no firm id at all	
A		Diff to rule 1: no street number	
A		Diff to rule 1: no postcode	
A	**8**	Diff to rule 1: no firm id at all	
A		Diff to rule 1: no street number	
A		Diff to rule 1: no street	
A	**9**	Diff to rule 1: no street	
A		Diff to rule 1: no postcode	
A		Diff to rule 1: no street number	
A	**10**	Diff to rule 1: firm id from patent data base	
A		Diff to rule 1: no street	
A		Diff to rule 1: no postcode	
A		Diff to rule 1: no street number	
A	**11**	Diff to rule 1: no firm id at all	
A		Diff to rule 1: no street	
A		Diff to rule 1: no postcode	
A		Diff to rule 1: no street number	
B	**12**	Last name	Last name
B		First name (only the first)	First name

B		Firm id found from registry	Firm id (**FIRST** employment period)
B		Application year	Registry year
B	**13**		Diff to rule 12: Firm id (**Second** employment period)
B	**14**		Diff to rule 12: Firm id (**Third** employment period)
B	**15**		Diff to rule 12: Firm id (**Fourth** employment period)
B	**16**		Diff to rule 12: Firm id (**Fifth** employment period)
B	**17**	Diff to rule 12: firm id from patent data base	Firm id (**FIRST** employment period)
B	**18**	Diff to rule 12: firm id from patent data base	Diff to rule 12: Firm id (**Second** employment period)
B	**19**	Diff to rule 12: firm id from patent data base	Diff to rule 12: Firm id (**Third** employment period)
B	**20**	Diff to rule 12: firm id from patent data base	Diff to rule 12: Firm id (**Fourth** employment period)
B	**21**	Diff to rule 12: firm id from patent data base	Diff to rule 12: Firm id (**Fifth** employment period)
B	**22**	Diff to rule 12: second firm id from patent data base	Firm id (**FIRST** employment period)
B		Diff to rule 12: postcode	Diff to rule 12: postcode
B	**23**	Diff to rule 12: second firm id from patent data base	Diff to rule 12: Firm id (**Second** employment period)
B		Diff to rule 12: postcode	Diff to rule 12: postcode
B	**24**	Diff to rule 12: second firm id from patent data base	Diff to rule 12: Firm id (**Third** employment period)
B		Diff to rule 12: postcode	Diff to rule 12: postcode
B	**25**	Diff to rule 12: second firm id from patent data base	Diff to rule 12: Firm id (**Fourth** employment period)
B		Diff to rule 12: postcode	Diff to rule 12: postcode
B	**26**	Diff to rule 12: second firm id from patent data base	Diff to rule 12: Firm id (**Fifth** employment period)
B		Diff to rule 12: postcode	Diff to rule 12: postcode
B	**27**	Diff to rule 12: second firm id from patent data base	Firm id (**FIRST** employment period)
B	**28**	Diff to rule 12: second firm id from patent data base	Diff to rule 12: Firm id (**Second** employment period)
B	**29**	Diff to rule 12: second firm id from patent data base	Diff to rule 12: Firm id (**Third** employment period)
B	**30**	Diff to rule 12: second firm id from patent data base	Diff to rule 12: Firm id (**Fourth** employment period)
B	**31**	Diff to rule 12: second firm id from patent data base	Diff to rule 12: Firm id (**Fifth** employment period)
B	**32**	Diff to rule 12: second firm id from patent data base	Firm id (registry, last week of the year)
C	**33**	Last name	Last name
C		Postcode	Postcode

C		Street	Street
C		First name (first part. E.g. "Karl-Erik" we use "Karl")	First name
C		Application year	Registry year
C	**34**		Diff to rule 33: Last name is previous last name
C	**35**		Diff to rule 33: Other first name
C	**36**	Diff to rule 33: (first part. E.g. "Karl-Erik" we use "Karl")	Diff to rule 33: Other first name (first part. E.g. "Karl-Erik" we use "Karl")
C	**37**	Diff to rule 33: First name as in rule 12	Diff to rule 33: Street name in Swedish
C	**38**	Diff to rule 33: First name (first part. E.g. "Karl-Erik" we use "Karl")	Diff to rule 33: Street name in Swedish
C	**39**	Diff to rule 33: First name (first part. E.g. "Karl-Erik" we use "Karl")	Diff to rule 33: Other first name
C			Diff to rule 33: Street name in Swedish
C	**40**	Last name	Last name
C		Postcode	Postcode
C		First name	First name
C		Application year	Registry year
C	**41**	Diff to rule 40: First name (first part. E.g. "Karl-Erik" we use "Karl")	
C	**42**	Diff to rule 40: First name (first part. E.g. "Karl-Erik" we use "Karl")	Diff to rule 33: Other first name
C	**43**	Diff to rule 40: First name (first part. E.g. "Karl-Erik" we use "Karl")	Diff to rule 33: First name, part one
C	**44**	Diff to rule 40: First name (first part. E.g. "Karl-Erik" we use "Karl")	Diff to rule 33: Other name, part one
D	**45**	Last name	Last name
D		First name	First name
D		Postcode	Postcode
D		Street	Street
D		Application year	Registry year t-1
D	**46**	Last name	Last name
D		First name	First name
D		Postcode	Postcode
D		Street	Street
D		Application year	Registry year t-2
D	**47**	Last name	Last name
D		First name	First name
D		Postcode	Postcode

D		Street	Street
D		Application year	Registry year t-3
D	**48**	Last name	Last name
D		First name	First name
D		Postcode	Postcode
D		Street	Street
D		Application year	Registry year t+1
D	**49**	Last name	Previous last name
D		First name	First name
D		Postcode	Postcode
D		Street	Street
D		Application year	Registry year t-1
D	**50**	Last name	Previous last name
D		First name	First name
D		Postcode	Postcode
D		Street	Street
D		Application year	Registry year t-2
D	**51**	Last name	Previous last name
D		First name	First name
D		Postcode	Postcode
D		Street	Street
D		Application year	Registry year t-3
D	**52**	Last name	Previous last name
D		First name	First name
D		Postcode	Postcode
D		Street	Street
D		Application year	Registry year t+1
D	**53**	Last name	Last name
D		First name	Other first name
D		Postcode	Postcode
D		Street	Street
D		Application year	Registry year t-1
D	**54**	Last name	Last name
D		First name	Other first name
D		Postcode	Postcode
D		Street	Street
D		Application year	Registry year t-2
D	**55**	Last name	Last name
D		First name	Other first name
D		Postcode	Postcode
D		Street	Street
D		Application year	Registry year t-3
D	**56**	Last name	Last name
D		First name	Other first name
D		Postcode	Postcode
D		Street	Street
D		Application year	Registry year t+1

D	**57**	Last name	Last name
D		First name	Other first name
D		Firm id from patent data base	Firm id (registry, last week of the year)
D		Application year	Registry year t-1
D	**58**	Last name	Last name
D		First name	Other first name
D		Firm id from patent data base	Firm id (registry, last week of the year)
D		Application year	Registry year t-2
D	**59**	Last name	Last name
D		First name	Other first name
D		Firm id from patent data base	Firm id (registry, last week of the year)
D		Application year	Registry year t-3
D	**60**	Last name	Last name
D		First name	Other first name
D		Firm id from patent data base	Firm id (registry, last week of the year)
D		Application year	Registry year t+1
D	**61**	Last name	Last name
D		First name	Other first name
D		Firm id found from registry	Firm id (registry, last week of the year)
D		Application year	Registry year t-1
D	**62**	Last name	Last name
D		First name	Other first name
D		Firm id found from registry	Firm id (registry, last week of the year)
D		Application year	Registry year t-2
D	**63**	Last name	Last name
D		First name	Other first name
D		Firm id found from registry	Firm id (registry, last week of the year)
D		Application year	Registry year t-3
D	**64**	Last name	Last name
D		First name	Other first name
D		Firm id found from registry	Firm id (registry, last week of the year)
D		Application year	Registry year t+1

2. Command file structure

The command files (Stata do-files) are structured as follows:

a. The do-file *returns_MASTER_DATAPREP_date.do* executes all the other do-files. The main do-files, and the auxiliary do-files they execute are the following:

   a. *returns_DATAPREP_1_CLEANED_date.do*

i. runs *Inventors_patstat_date.do*. This file reads in *patstat_19782013.dta* and transforms it into a format where the observation unit is an individual-year observations (instead of individual-patent). It creates file *patstat_all.dta*.
ii. reads in registry data *fleed_kokonais_`i'.dta* which each contain all the working age residents in Finland in a given year (`i' refers to calendar year). Saves needed variables and creates an individual-year panel to which *patstat_all.dta* is merged (by individual and year).
iii. Creates data set *returns_DATAPREP_1_CLEANED.dta*.

b. *Returns_DATAPREP_2_date.do*.
i. Reads in *returns_DATAPREP_1_CLEANED.dta*.
ii. Matching is done by socioeconomic class. The socioeconomic classes used are: senior management, senior white-collar worker, junior management, junior white-collar worker.
iii. Generates dataset with the name structure *returns_EST_`j'G.dta* which are used to form the estimation sample and j stands for a socioeconomic class.

b. The do-file *Returns_MASTER_ESTIMATION_date.do* executes the estimations. These are broken down by estimation sample, i.e., there is a separate do-file for each, but the codes are otherwise identical.
a. Returns_EST_CEM_cluster_date.do
b. Returns_EST_CEM_firstinv_date.do
c. Returns_EST_CEM_noPhD_date.do
d. Returns_EST_CEM_stem_date.do
e. Returns_EST_CEM_nostem_date.do
f. Returns_EST_CEM_ownfirm_date.do

3. Matching the individual characteristics data from Statistics Finland with the inventor data.

This is achieved using the encrypted individual identifiers in the registry data, and the year of patent application from PATSTAT.

4. Forming the matched estimation sample

The estimation sample is constructed in two steps. First, the socio-economic status variables are generated in *returns_DATAPREP_1_CLEANED_date.do*. Second, in preparation for the CEM matching, the matching variables are generated in do-file *returns_DATAPREP_2_date.do*.

Step 1: The socio-economic variables are based on the registry variable *sose*. We impute missing values backward and forward by up to seven years.

sose_mngt_sr	dummy variable taking value 1 if two first digits of *sose* are 31 and 0 otherwise
sose_sr	dummy variable taking value 1 if first digit of *sose* is 3 and the first two digits are not 31, and 0 otherwise
sose_mngt_jr	dummy variable taking value 1 if two first digits of *sose* are 41 and 0 otherwise

sose_jr dummy variable taking value 1 if first digit of *sose* is 4 and the first two digits are not 41, and 0 otherwise

sose_bc dummy variable taking value 1 first digit of *sose* is 5 and 0 otherwise.

Step 2: The matching variables are defined as

inventor_all$_{it}$ 0 = individual i never invents; 1 = individual i invents at least once; 2 = individual i invents in year t.

msc 0 = education level less than MSc; 1 = education level at least MSc, i.e., first digit of registry variable *ututku* is 5 or higher

edu_science 0 = education field not STEM; 1 = second digit of registry variable *ututku* is 4 (technology and science), 5 (transport and computer science) or 7 (agriculture and forestry).

sme 0 otherwise, 1 = number of employees less than 250. Number of employees of firm j in year t is generated by calculating how many individuals are registered at firm j in year t.

firm_size_q quartile of firm size distribution. Generated year by year using the number of employees. Number of employees of firm j in year t is generated by calculating how many individuals are registered at firm j in year t.

manuf The industry classification used by Statistics Finland changes over time. We follow the convention and define the following 2-digit industries to be in manufacturing:

TOL code 11 – 29 up to 1992

TOL code 15 – 39 1993-2000

TOL code 15- 37 2001 – 2006

TOL code 10 – 33 2007-

region_large 1 = individual i in Southern or Western Finland in year t; otherwise 2. Based on registry variable *suuralue*.

age 4 age groups: at most 30; 31-40; 41-50, over 50. Based on registry variable *ika* (age in years)

dtf time since last degree (distance to the human capital frontier). Calculated as the difference between year (registry variable *vuosi*) and the year of obtaining last degree (registry variable *suorv*)

Matching is done using the Stata cem – command (see cem: Coarsened exact matching in Stata Blackwell, Iacus, King and Porro, The Stata Journal 9, Number 4, pp. 524–546). The matching is done by socioeconomic group and year, without replacement. The matching vector consists of the following variables: *MSc*, *edu_sciene*, *manuf*, *firm_size_q*, *region_large*, *age*, *dtf*, *suorv_d*. Matching is exact otherwise, but *dtf* is matched in bins of 5 years (0-5, 6-10, 11-15, 16-20, 21-).

We use k2k option, which means that we end up having the same number of treated and control units (within all strata defined by the matching variables) in the estimation sample. As is explained Blackwell, Iacus, King and Porro (2009), the k2k pruning occurs within a stratum by random matching inside cem strata.

5. Descriptive statistics analysis

The descriptive statistics reported in Appendix A2 are produced by *returns_CEM_cluster_date.do*.

6. Regression (DID) analysis.

Dependent variables are defined in the estimation do-files as

lnwage natural log of deflated wage (deflator is the Finnish consumer price index taken from table *030_khi_tau.xlsx* of Statistics Finland). Wage is calculated as the sum of registry variables *tyrtu* (entrepreneurial income) and *tyotu* (wage income).

ue_d dummy-variable based on registry variable *ptoim1*: *ptoim1* = 11 if individual i is employed in year t. Takes value 0 if *ptoim1* = 11 and 1 otherwise.

Explanatory variables are defined as follows:

msc see above

dtf see above

BSc dummy taking value 1 if the first digit of the registry variable *ututku* is 5, 6, 7, 8; 0 otherwise.

Senior dummy variable taking value 1 if individual i is at least 40 in year t and 0 otherwise

Suorv_d dummy taking value 1 if registry variable *suorv* (year of last degree) not observed for individual i in year t.

How to read the regression tables?

The estimation code produces regression tables both directly into the log – file and into excel tables. To facilitate reading these tables, we below list the treatment variables of interest.

The treatment effect variables are denoted *treated_pre* and *treated_post*. The interactions between these and the different measures of observed heterogeneity (senior, MSc/BSc, DTHCF = distance to the human capital frontier)) are denoted

Treated_pre_ia2 interaction between the dummy treated_pre and the dummy for high education (MSc for white-collar workers, BSc for blue-collar workers)

Treated_pre_ia3 interaction between the dummy treated_pre and DTHCF (distance to the human capital frontier)

Treated_pre_iasenior interaction between the dummy treated_pre and the dummy for being a senior

Treated_pre_iasuorv_d interaction between the dummy treated_pre and the dummy for the year of last degree missing.

Treated_post_ia2	interaction between the dummy treated_post and the dummy for high education (MSc for white-collar workers, BSc for blue-collar workers)
Treated_ post _ia3	interaction between the dummy treated_post and DTHCF distance to the human capital frontier)
Treated_post_iasenior	interaction between the dummy treated_post and the dummy for being a senior
Treated_post_iasuorv_d	interaction between the dummy treated_post and the dummy for the year of last degree missing.