FORTE 2021-2027
Matching the test data set

Laurent Bergé

January 13, 2023

Abstract

This document reports descriptive statistics on the matching of inventors to Swedish persons using the test data set.

Contents

1 Input data

This section presents the input data and their main descriptive statistics.

1.1 Statistics Sweden (Stat-SE) data

There are two main data sources:

- individual names, birth date and private addresses (**OE_lev_RTB**)
- individual employer and the address of that employer (**OE_lev_Jobb**)

These two data sets are available yearly. However, their year availability differ:

Data set	Start year	End year	Variables of interest
RTB	1978	2020	individual name, birth date, address
Jobb	1985	2020	employer name, employer address

The two data sets are linked via an individual identifier (**LopNr**).

1.1.1 Descriptive statistics

Note that since this is a test data set, we dispose of only a select sample of these two data sets.

The data in RTB reports the names and addresses of 1,373,780 unique individuals, leading to 10,045,153 unique identities x addresses, whereby the names and/or addresses may change for each individual.

1.2 Patent data

The data source is REGPAT from the OECD, version February-2022 (Maraut et al., 2008). It covers all patents filed at the EPO from 1977 onward and reports (among other) the addresses of both the applicant and the inventors.

There are three data sets we are interested in, they are detailed below:

Data set	Variable	Meaning
EPO_INV_REG		
	Appln_id	patent application identifier
	Inv_name	raw inventor name
	Address	raw inventor address
EPO_APP_REG		
	Appln_id	patent application identifier
	App_name	raw applicant name
	Address	raw applicant address
EPO_IPC		
	Appln_id	patent application identifier
	Prio_year	priority year (first filing)

1.2.1 Descriptive statistics

We restrict the sample to inventors reporting an address located in Sweden. This leads to 85,359 unique patents and 171,220 patent-inventors. The objective of the full algorithm is to match these 171,220 patent-inventors.

1.3 Terminology

Across the data sets created by the algorithm, there are two keys that are extensively used:

- id_se: refers to the Stat-SE person identifier, which uniquely identifies people residing in Sweden (the variable LopNr)
- id_inv_seq: refers to the patent-inventor identifier

Throughout this document, to avoid ambiguity, these variable names are used as regular words.

2 General information on cities and names

The following figures report back-to-back comparisons of the distributions of names and cities in patent and Stat-SE data sets:

- Figure 1: city comparison
- Figure 2: first names comparison
- Figure 3 : family names comparison

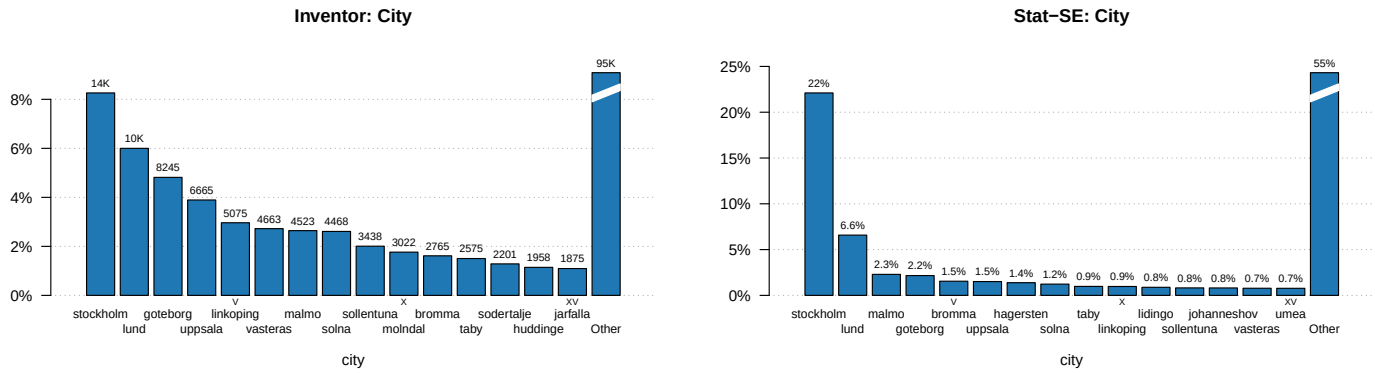

Figure 1: Comparison of the distribution of the cities of residence for the patent and STAT-SE data sets.

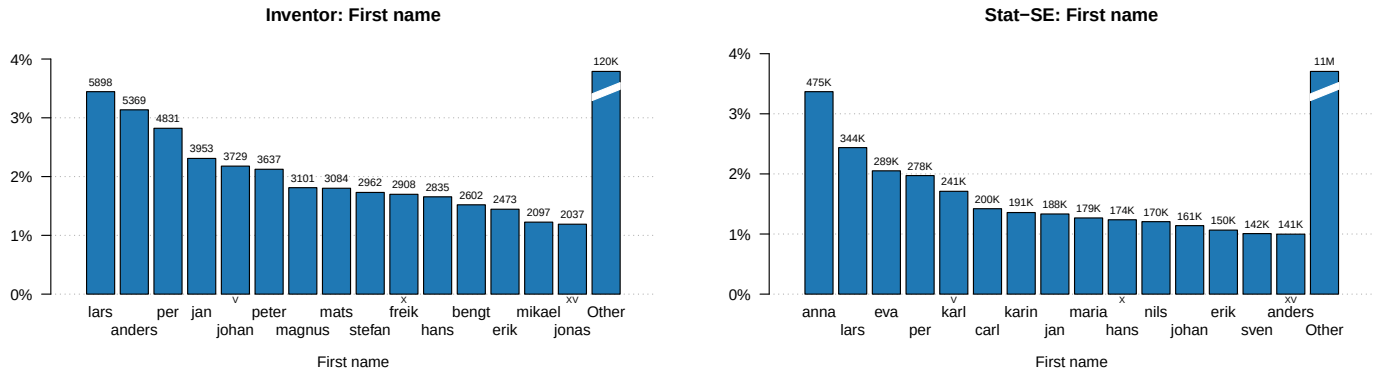

Figure 2: Comparison of the distribution of the first names for the patent and STAT-SE data sets.

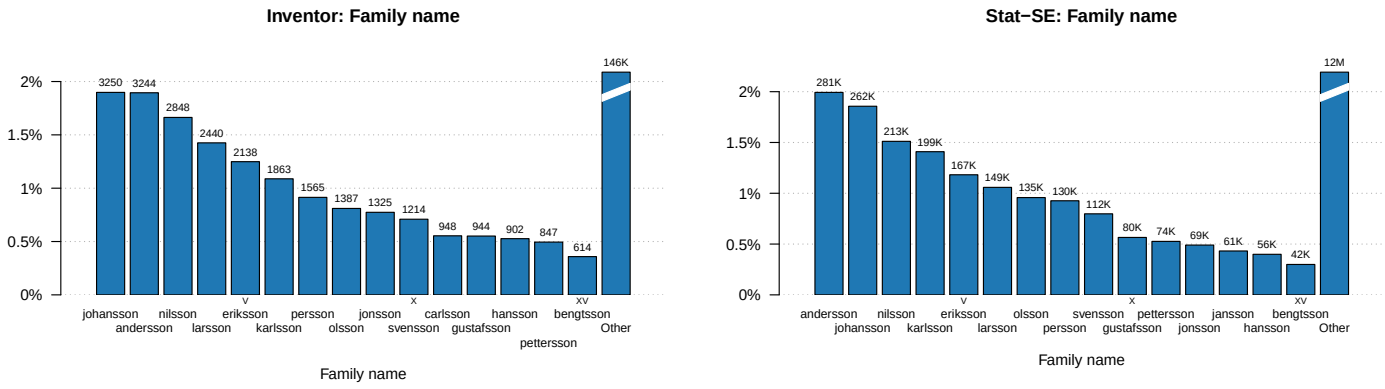

Figure 3: Comparison of the distribution of the family names for the patent and STAT-SE data sets.

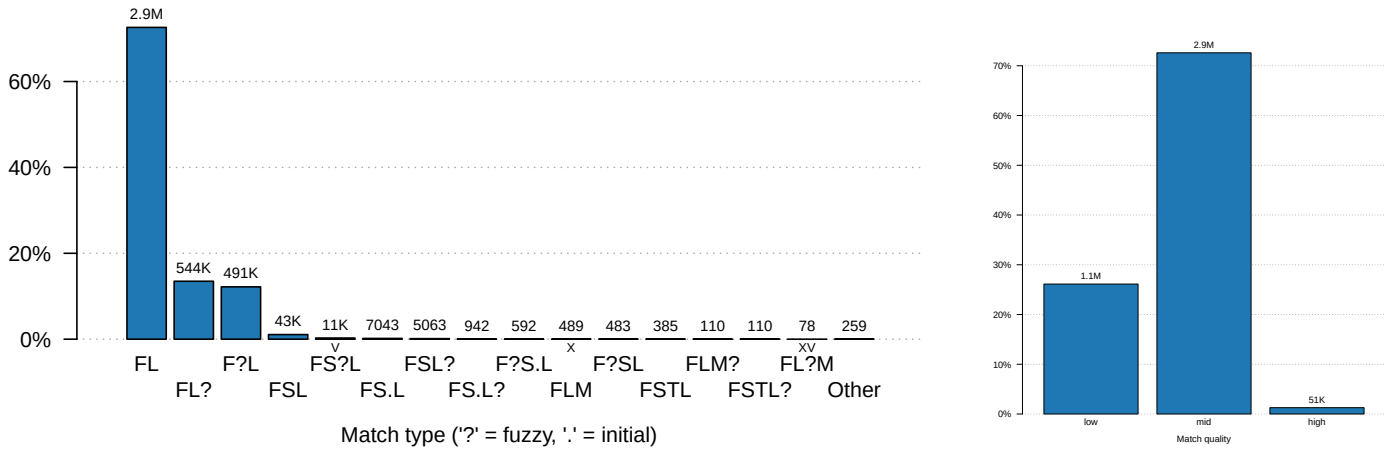

Figure 4: Match type and quality when matching by names. F means first first name, S (resp. T) second (resp. third) first name. L means last name, M means maiden name (in fact it is the second family name).

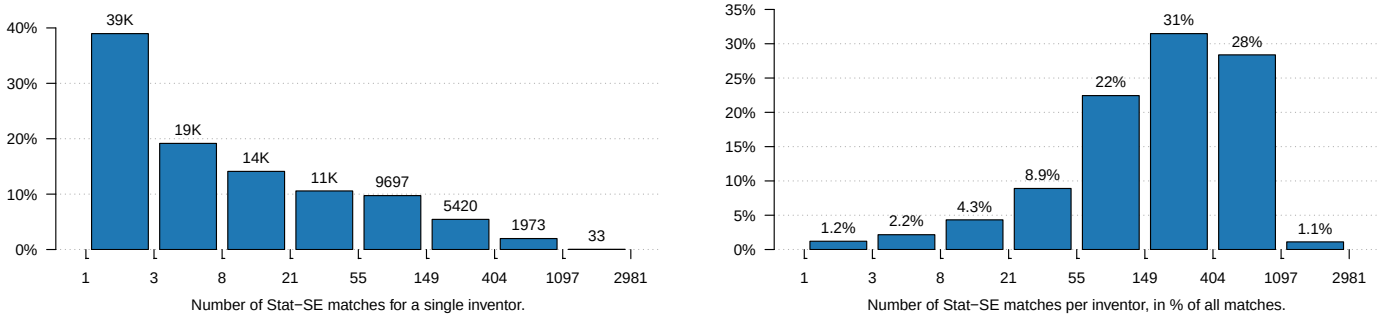

Figure 5: Distribution of the number of matches per inventor.

3 Matching procedure

3.1 Step 1: Matching by names

The idea of the name matching is to always take the largest set of information available to compare two persons. For instance Carl Friedrich Gauss will never be the same as Carl Frank Gauss but could be the same as Carl F. Gauss.

The matching is performed using up to three first names and two family names.

Result. In this first step we end up with 4,032,816 $id\_se \times id\_inv\_seq$ pairs. it corresponds to 99,617 inventor-patents and to 263,968 Stat-SE unique persons.

The matching is rather weak in general, being performed on the first first name and the family name in most cases, as reported in Figure 4. This may come down to the fact that inventors may simply report only their first first name.

The distribution of the matches is highly skewed towards names that are extremely common. This is reported in Figure 5 in which we can see that less than 4% of the matching pairs have less than 7 namesakes. Over 80% of the sample refer to matching pairs with more than 55 namesakes.

3.2 Step 2: Bilateral variables

For each $id\_se \times id\_inv\_seq$ pair found in the first step (the *potential* matches), we create the variables that will be used to inform the matching.

3.2.1 Same address

This variable represents whether a $id\_se \times id\_inv\_seq$ pair share the same address. There are three types of cases :

- *unmatched*: the addresses don't match

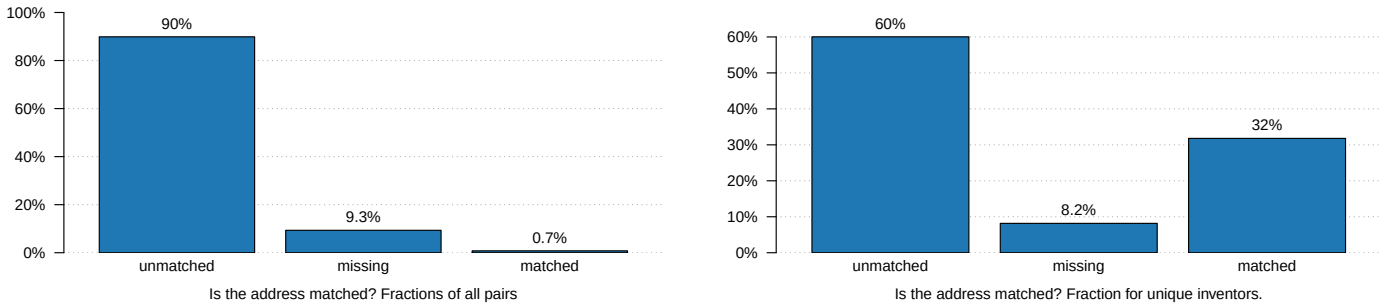

Figure 6: Distribution of the type of address matching, across all $id\_se \times id\_inv\_seq$ pairs (4M obs.) or for each id_inv_seq, i.e. inventor-patent (99k obs.).

- *matched*: the addresses match
- *missing*: the inventor's address is the one of the employer

The distribution of the types of is reported in Figure 6. We can see that, unsurprisingly, most pairs have addresses that do not match (only 0.7% match). However, 32% of inventor-patent have an address that matches and 8.2% do not report an address.

3.2.2 Same employer

This binary variable represents whether a $id\_se \times id\_inv\_seq$ pair shares the same employer. It is based on the name of the employer, and the address of the employer is also leveraged to find more matches. The results are reported in this table:

	Has same employer	
Type	%	Number
all $id\_se \times id\_inv\_seq$ pairs	1.99%	80,317
unique inventor-patent (id_inv_seq)	35.8%	35,621

Across all $id\_se \times id\_inv\_seq$ pairs, about 2% have the same employer. If we look at the share of inventors who is matched to at least one id_se sharing his employer, this number reaches 36%.

3.2.3 Other variables

We create a name probability that we combine with the name matching quality variable (e.g. fuzzy matching = low quality; two full first names = high quality). We will use them in the EM algorithm. The idea is that the probability to be a good match should be higher, ceteris paribus, as the name probability goes down (rare names are more likely to refer to the same person) or the match quality goes up (pairs whose first and second first-names as well as the first and second family names match are more likely to refer to the same person).

3.3 Step 3: EM algorithm & regular matching

3.3.1 EM matching

By construction, the number of $id\_se \times id\_inv\_seq$ which are true matches is low (see Figure 5, 80% of the pairs correspond to a single id_inv_seq matched to more than 55 id_se). Indeed, there are 4M $id\_se \times id\_inv\_seq$ pairs and 100k unique id_inv_seq, this leads to *at most* 2.5% of true matches (since id_inv_seq, inventor-patents, should be matched to a single id_se, Stat-SE identity). This very low number leads the EM algorithm to not work well because the group of the true matches would contain too few observations.This leads the algorithm to create groups that are coherent but do not represent the group of true matches (it is coherent on other characteristics).

To make the EM algorithm run smoothly, I keep only the inventor-patents (id_inv_seq) who have 5 or less namesakes (id_se). This represents 53,714 inventor-patents (more than 60% of all inventor-patents).

To initialize the algorithm, I allocate all the pairs with a matched address to the same class. The end distributions found by the algorithm are reported in Figure 7.

Interpreting Figure 7. If we focus on Class 2, which is the class of the true matches, we can see that 86% of the $id\_se \times id\_inv\_seq$ pairs in that class have an address that matches. This means that 14% of the pairs of that class have been matched without having an address in common (that is, the proximity of the other variables compensated the absence of address in common). 73% of the pairs of Class 2 share the same employer and a high fraction of the pairs have a low name probability and rather high name matching quality.

On the other hand, the ones of Class 1 (the wrong matches) have a very dispersed age distribution, only 0.4% have a matching address and 1.3% share the same employer. In general, they have names with high probability and low name matching quality.

Overall, this is a very coherent picture.

Outcome. In total, 25% of the pairs are found to be of Class 2 (the class of the true matched). The EM algorithm also reports the probability for the observation to belong to each class. This probability can be used to infer a level of confidence in the matching.

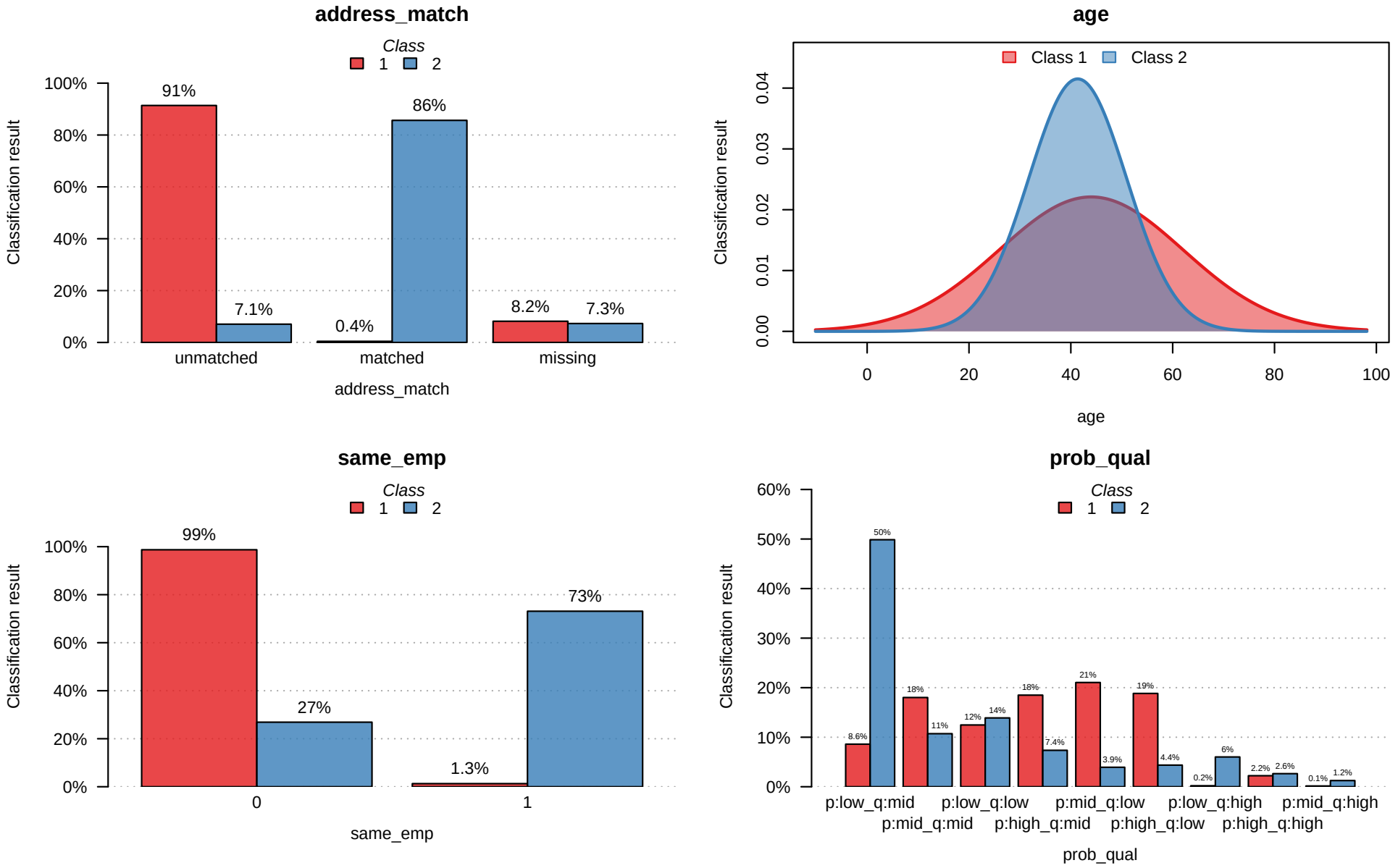

Figure 7: Meta distributions of the classes found with the EM algorithm.

Definitions: `prob_qual`: code giving first the name probability and second the name matching quality (e.g. fuzzy matching = low).

Using a high level of confidence of 80% (i.e. the probability to be a true match should be higher than 80%), we end up with 23,446 inventor-patents matched.

3.3.2 Descriptive information on the EM matching

Table 1 reports, a few $id\_se \times id\_inv\_seq$ pairs for which an id_inv_seq has been matched. For each id_inv_seq, we gather all the potential id_se as well as all the relevant information to judge the quality of the match (raw names, addresses, employers).

Table 2 is similar but reports only id_inv_seq who have been matched without any matching address.

Table 1: Descriptive information for the matched/unmatched $id\_se \times id\_inv\_seq$ pairs.

Definitions: `prob`: probability to be a true match; `prob_qual`: code giving first the name probability and second the name matching quality (e.g. fuzzy matching = low).

id_se	id_inv_seq	prob	address_match	same_emp	age	prob_qual	inv_name	se_name	inv_address	se_address	inv_emp	se_emp
1,152,716	5	0.99	matched	1	49	p:low, q:mid	fertner, antoni	antoni, fertner	kruthusbacken 76,s-169 52 solna	tre liljor 3, 3 1/2, 11344, stockholm; hagag 17 m, iii, 11347, stockholm; observatorieg 12, iii, ...	telefonab lm ericsson	stiftelsen institutet for mikrovagsteknik; stiftelsen institutet for mikroelektronik; stancord ab...
571,147	5	0.01	unmatched	0	20	p:low, q:mid	fertner, antoni	antoni julian, fertner	kruthusbacken 76,s-169 52 solna	tre liljor 3, 3 1/2, 11344, stockholm; hagag 17 m, iii, 11347, stockholm; observatorieg 12, iii, ...	telefonab lm ericsson	
1,129,489	7	0.99	matched	1	25	p:low, q:mid	braun, christian	christian, braun	lokstallsgatan 6,s-113 21 stockholm	tralargr 9, 17547, jarfalla; sandhamnsg 36 /von braun/, 11528, stockholm; furusundsg 2, 2 tr, 115...	allgon ab	vag center ab; kista bilcentrum ab; ahlens restauranter ab; avitec ab; allgon mobile communicatio...
364,584	10	0.99	matched	1	28	p:high, q:mid	ericson, petter	eric petter vilhelm, eriksson; eric petter vilhelm, ericson	industrigatan 2 b,s-212 14 malmo	nyponv 81, 93100, skelleftea; nyponv 81, 93149, skelleftea; snostigen 11, 93151, skelleftea; slat...	anoto ab	skelleftea kommun; skelleftea arbetarkommun; studieforbundet medborgarskolan; boliden mineral ab;...
258,854	10	0.01	unmatched	0	35	p:low, q:low	ericson, petter	jan peter, kanflo; john peter, ericson	industrigatan 2 b,s-212 14 malmo	bybergsv 12, 68300, hagfors; gardesv 6, 43080, hovas; molndalsv 3, 41263, goteborg; hantverkarg 5...	anoto ab	utrikesdepartementet; systembolaget ab; dagens nyheter ab; regeringskansliet; utkiken ord bilder ...
329,502	10	0	unmatched	0	38	p:high, q:low	ericson, petter	stig peter, eriksson; stig peter, ericson	industrigatan 2 b,s-212 14 malmo	slanbarsv 7, vi, 18236, danderyd; rindog 8, v, 11536, stockholm; folkungagatan 100, i og, 11630, ...	anoto ab	ansvar omsesidig sakforsakring for helnyktra; arla ekonomisk forening; emc computer systems ab; k...
680,801	10	0	unmatched	0	40	p:mid, q:low	ericson, petter	dan peter, ericsson; dan peter, ericson	industrigatan 2 b,s-212 14 malmo	verkstadsg 12, 78300, sater; hagerstensv 158, ii, 12653, hagersten; gransatragr 18, iii, 12736, s...	anoto ab	stockholms lans landsting; tjanstemannens centralorganisation tco; wiksten dan georg; taxi trafik...
417,212	10	0	unmatched	0	32	p:high, q:low	ericson, petter	peter yngve, eriksson; peter yngve, ericson	industrigatan 2 b,s-212 14 malmo	blabarsstigen 17, 79200, mora; norebergsv 10 c, 79200, mora; mortg 12, iii /telo/, 13343, saltsjo...	anoto ab	rikspolisstyrelsen; skogsstyrelsen; mora hotell ab; svenska mc donald s ab; aklagarmyndigheten; p...
632,198	30	0.99	matched	0	44	p:low, q:mid	mendelsohn, stuart	stuart michael, mendelsohn	solberga angsvag 15,s 125 44 stockholm	solberga angsvag 15, 12544, alvsjo; morellstigen 32 lgh 08, 93170, skelleftea; solberga angsvag 1...	nokia corporation	apis technical training ab; botkyrka kommun; protegrity research development ab; sigma kudos swed...
393,996	37	0.99	matched	1	25	p:high, q:mid	lundberg, ivan	ivan, lundberg	ibsengatan 80,168 47 bromma	m bagares gr 40 /r hannah/, 12355, farsta; lasarettsg 37 c, 57400, vetlanda; lasarettsg 37c, 5740...	abb ab	vasterleds direktreklam kommanditbolag; jaka reklamdistribution ab; svenska kabel tv ab; ericsson...
769,419	37	0	unmatched	0	46	p:high, q:mid	lundberg, ivan	ivan peter, lundberg	ibsengatan 80,168 47 bromma	svenshogsv 19, 22241, lund; vasterg 13, 22229, lund; castors v 5, 24500, staffanstorp; konsertv 6...	abb ab	akermans verkstad i lund ab; lundgrens mek verkstad ab; lundgren machinery ab; tomb packaging ab
175,551	37	0	unmatched	0	35	p:high, q:low	lundberg, ivan	ivars, krapas; ivar, krapas; ivar, lundberg krapas; ivar, lundberg	ibsengatan 80,168 47 bromma	kocksg 11, 4 tr /lundstrom/, 11624, stockholm; tallbacken 2, 95731, overtornea; kocksg 11, 4 tr, ...	abb ab	overtornea kommun; overtornea abf avd nr 201; abf nedre tornedalen; studieforbundet vuxenskolan o...
137,302	37	0	unmatched	0	88	p:high, q:mid	lundberg, ivan	ivan harry, lundberg	ibsengatan 80,168 47 bromma	kjolerod, 44060, skarhamn; molnedal 802, 47192, klovedal	abb ab	
1,064,611	37	0	unmatched	0	92	p:high, q:low	lundberg, ivan	ivar, lundberg	ibsengatan 80,168 47 bromma	gullaker, 51051, ganghester; gullaker, 50278, ganghester; gullaker, 50771, ganghester	abb ab	

Table 2: Results of the algorithm. Only unmatched addresses.

Definitions: `prob`: probability to be a true match; `prob_qual`: code giving first the name probability and second the name matching quality (e.g. fuzzy matching = low).

id_se	id_inv_seq	prob	address_match	same_emp	age	prob_qual	inv_name	se_name	inv_address	se_address	inv_emp	se_emp
1,089,090	97	0.98	unmatched	1	30	p:low, q:high	herrero verÃ³n, christian	christian, herrero veron	siporexgatan 93,s-240 10 dalby	fagottgranden 23 a 259, 22468, lund; nationsgatan 5 2, 22460, lund; skyttelinjen 120, 22649, lund	telefonab lm ericsson	ericsson mobile communications ab; ericsson mobile platforms ab; ericsson ab; huawei technologies...
687,450	143	0.93	unmatched	1	35	p:low, q:mid	rotticci, didier	didier henri, rotticci	astrazeneca rd sodertalje,s-151 85 sodertalje	kungshamra 31, 113, 17070, solna; bastug 19, 5 tr /orrenius/, 11825, stockholm; svanholmsvagen 2 ...	astrazeneca ab	tekniska hogskolan i stockholm; kungliga tekniska hogskolan; astrazeneca ab
1,155,977	153	0.99	missing	1	37	p:low, q:mid	svenmar, peter	peter, svenmar	c/o skanska sverige ab skanska stomsystem box 35,s-245 02 hjarup	langg 137, 58267, linkoping; magle lilla kyrkogata 12, 22351, lund; magle lilla kyrkogata 8, 2235...	skanska sverige ab	linkopings segelsallskap; ostergotlands elektriska ab; johansson ulf karl gustav; hjulsbro spanna...
1,362,113	223	0.75	missing	1	62	p:mid, q:mid	olsson, bozena nenna	bozena, olsson; nenna bozena, olsson	p.o. box 88,246 21 loddekopinge	s ljungv postl 202, 24402, furulund; magneg 16, 24402, furulund; postl 374, 24021, loddekopinge; ...	olsson bozena nenna	hemmets journal ab; arbetsmarknadsverket; stiftelsen tjanstemannens trygghetsfond; kma kemimaklar...
1,051,451	229	0.93	unmatched	1	37	p:low, q:mid	svendenius, jacob	carl jacob, svendenius	margaretavagen 31,22240 lund	overlararev 2 b, 22367, lund; griffelv 10, 22367, lund; griffelv 10, 22467, lund; sodra esplanade...	volvo car corporation; haldex brake products ab	jms i reklamgarden ab; stiftelsen tem vid lunds universitet; garantiforeningen for folkhogskolan ...
867,159	437	0.93	unmatched	1	37	p:low, q:mid	broddefalk, johan	johan olof, broddefalk	astrazeneca rd molndal,s-431 83 molndal	jatteg 83, 59300, vastervik; getingstigen 18, 59300, vastervik; getingstigen 18, 59352, vastervik...	astrazeneca ab	slipnaxos ab; kalmar lans landsting; becker pulver system ab; electrolux cleaning appliances ab; ...
1,084,014	480	0.62	unmatched	1	41	p:mid, q:mid	winter, ulf	ulf joseph, marklund; ulf joseph, winter	laby-osterby 4:20,75326, uppsala	norrbackag 1 b, 93100, skelleftea; norrbackag 1 b, 93137, skelleftea; swedenborgsg 34 a /finnson/...	q med ab	sara hotels ab; boliden metall ab; norra vasterbotten tidningsab; skelleftea video recordline ab;...
161,037	480	0.03	unmatched	0	26	p:low, q:mid	winter, ulf	, winter; ulf bengt andreas, winter	laby-osterby 4:20,75326, uppsala	kantelev 12, 19635, kungsangen; kantelev 12, 19637, kungsangen; kantelevagen 12, 19637, kungsange...	q med ab	swedish neutral ab; winter brothers ab
1,043,193	538	0.93	unmatched	1	45	p:low, q:mid	tapper, paul	paul michael, tapper	spexhults herrgard,57195 nassjo	farmgr 3, 25222, helsingborg; o d krooks g 72, 25243, helsingborg; vildandsv 24 c,:301, 22234, lu...	tapper paul	universitetet i lund; lunds kommun; malmohus lans landsting; handelsbolag filmhaftet; bokforlaget...
168,039	581	0.8	unmatched	1	59	p:low, q:mid	nicander, ingrid	ingrid anna elisabeth, nicander; ingrid anna e, wÃ¥rdell nicander; ingrid anna elisabeth, nicande...	platavagen 6,s-136 73 huddinge	platav 6, 13671, handen; platav 6, 13672, haninge; raggatan 12 2 tr, 11859, stockholm; platavagen...	scibase ab	stockholms kommun; finax finans service ab; finax finans ab; haninge kommun; nova park paravan ho...
872,042	582	0.87	unmatched	1	26	p:low, q:mid	birgersson, ulrik	hans ulrik birgersson, nissfolk; hans ulrik, birgersson	timotejgatan 4,s-118 59 stockholm	ljusstoparbacken 22 b, ii, 11745, stockholm; ljusstoparbacken 22 b, i, 11745, stockholm; ljusstop...	scibase ab	mediadirekt i osteraker ab; fundraising gruppen i stockholm ab; constantine joe; kungliga teknisk...
618,191	582	0	unmatched	0	34	p:high, q:low	birgersson, ulrik	eva ulrika, birgersson; eva ulrika, lorenzen	timotejgatan 4,s-118 59 stockholm	nyponv 166, 61300, oxelosund; stagneliusg 25, 39234, kalmar; nyponv 166, 61338, oxelosund; backas...	scibase ab	oxelosunds forsamling; oxelosunds kommun; ssab svenskt stal ab; varuhallen prisman i oxelosund ab...
722,478	718	0.94	unmatched	1	40	p:low, q:mid	willars, per	per hans Ã¥ke, willars	ytterbystrand 6,s-185 94 vaxholm	kungsg 33, 18500, vaxholm; grev tureg 53, nb, 11438, stockholm; furusundsg 11,6 tr, 11537, stockh...	telefonab lm ericsson	stockholms lans landsting; tbv solna sundbyberg; recognition equipment industri ab; tekniska hogs...

3.3.3 Regular matching

For the inventor-patents which have been matched to 6 or more id_se (Swedish identities), we matched them simply using the information on the address. There were 14,012 such matches.

3.3.4 Summing up

At the end of this step, we have matched 37,458 inventor-patents to single Swedish identities (it represents 20% of *all* the inventor-patents).

3.4 Step 4: Adding patent information

We use the patents that were identified in the previous step to construct the patent pools of each id_se (inventor-patent). Since all matched id_se (Swedish identity) now have patent information attached to them, we use this information to create patent-related distances with inventor-patents that were not matched.

There are 41,320 unique patents that were either produced by matched id_se or to which a matched id_se is a potential match[1] (but was not matched in the previous step). These patents are distributed in 37,458 matched and 20,157 unmatched inventor-patents (it is normal that the sum is higher than 41,320). The objective here is to leverage patent information to match the 20,157 unmatched inventor-patents.

3.4.1 Bilateral patent related variables

We construct 3 patent related variables:

- same coauthor
- same technology
- same applicant

For each variable, there is a comparison between the characteristics of the inventor-patent with the patent pool of the Swedish identity which is a potential match (i.e. it has been matched by name only).

Same co-author. This variable is binary and equal to 1 if the inventor-patent and the pool of patents of the Swedish identity share at least one co-author. The co-author is identified using the first name and the last name, of course self is excluded in the comparison.

We end up with 40% of the pairs having a common co-author.

Technological proximity. To measure technological proximity, we make use of IPC codes. IPC codes take the form of a character string, for instance "C05G001/02" which represents the class "Mixtures of fertilisers covered individually by different subclasses of class of superphosphates with ammonium nitrate". As we can see, this is very specific, and if we use only full IPC codes to compare two patents, one would get very few matches, even for patents written by the same person.

Instead of using the full IPC code, we will use the IPC at 1, 3, 4 and 7 digits which represent different levels of precision. For example, we can break down "C05G001/02" as:

Digits	IPC	Meaning
1	C	chemistry; metallurgy
3	C05	fertilisers; manufacture thereof
4	C05G	mixtures of fertilisers covered individually by different subclasses of class c05; mixtures of one or more fertilisers with additives not having a specific fertilising activity; fertilisers characterised by their form
7	C05G001	mixtures of fertilisers covered individually by different subclasses of class ...
9	C05G001/02	mixtures of fertilisers covered individually by different subclasses of class of superphosphates with ammonium nitrate

Hence we create a categorical variable whose value is the highest digit at which an inventor-patent matches the IPCs of the patent pool of the Swedish identity. For instance if an inventor-patent has an IPC of "C22C 5/08" and in the patent pool of the Swedish identity, all patents are in IPC "C" and the closest IPC is "C22B 7/02", then the variable will take value 3.

If there are no IPC matches, even at the 1-digit level, then the variable will take value 0. The distribution of IPC matches is reported in Figure 8. We see that the majority of $id\_se \times id\_inv\_seq$ pairs, 62%, do not have a single IPC in common, even at the broadest level. About 6% of the pairs have patents sharing a 7-digits IPC code.

[1] A potential match is an $id\_se \times id\_inv\_seq$ pair which has been matched by name in Step 1: Matching by names.

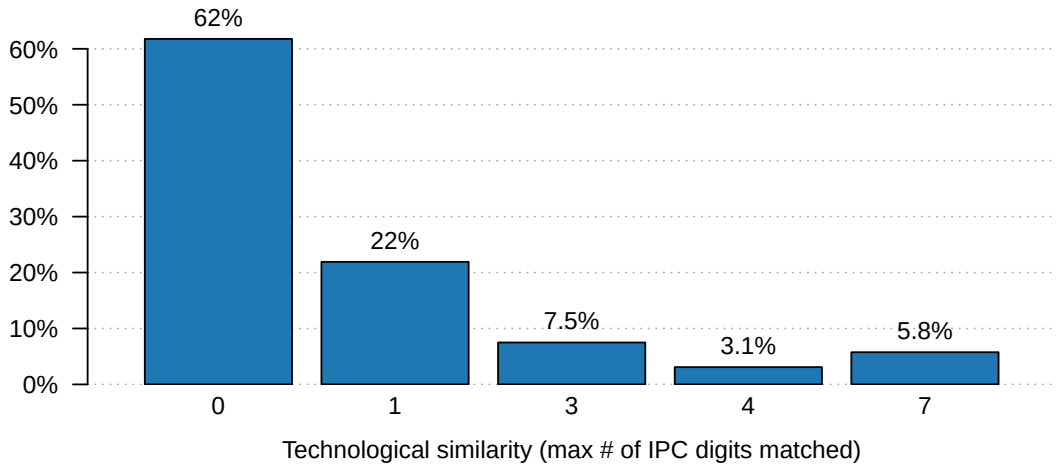

Figure 8: Technological similarity between inventor-patents (id_inv_seq) and the patent pools of Swedish identities (id_se).

Same applicant. For large companies with multiple entities, it is common to produce the invention in a firm but set another firm as the applicant of the patent (e.g. the one which manages intellectual property). Hence this will lead to mismatches between the employer of the Swedish person and the applicant of the patent. To avoid this we look directly if the applicant of an inventor-patent matches at least one applicant of the pool of patents of the Swedish identity.

We end up with a binary variable taking value 1 for the pairs having at least one common applicant. We end up with 5.8% of the pairs having a common applicant.

3.4.2 EM matching with patent data

Once we have created the previous variables, we add them to the existing distance variables and apply the EM algorithm to the 68,781 unmatched inventor-patents × Swedish identity pairs (it represents 20,157 unique inventor-patents and 3,949 unique Swedish identities).

The algorithm finds 7.8% of the pairs to be true matches. The distributions of the parameters of the variables used to compute the match probability are in Figure 9.

We end up with 4,203 new inventor-patents being matched to Swedish identities.

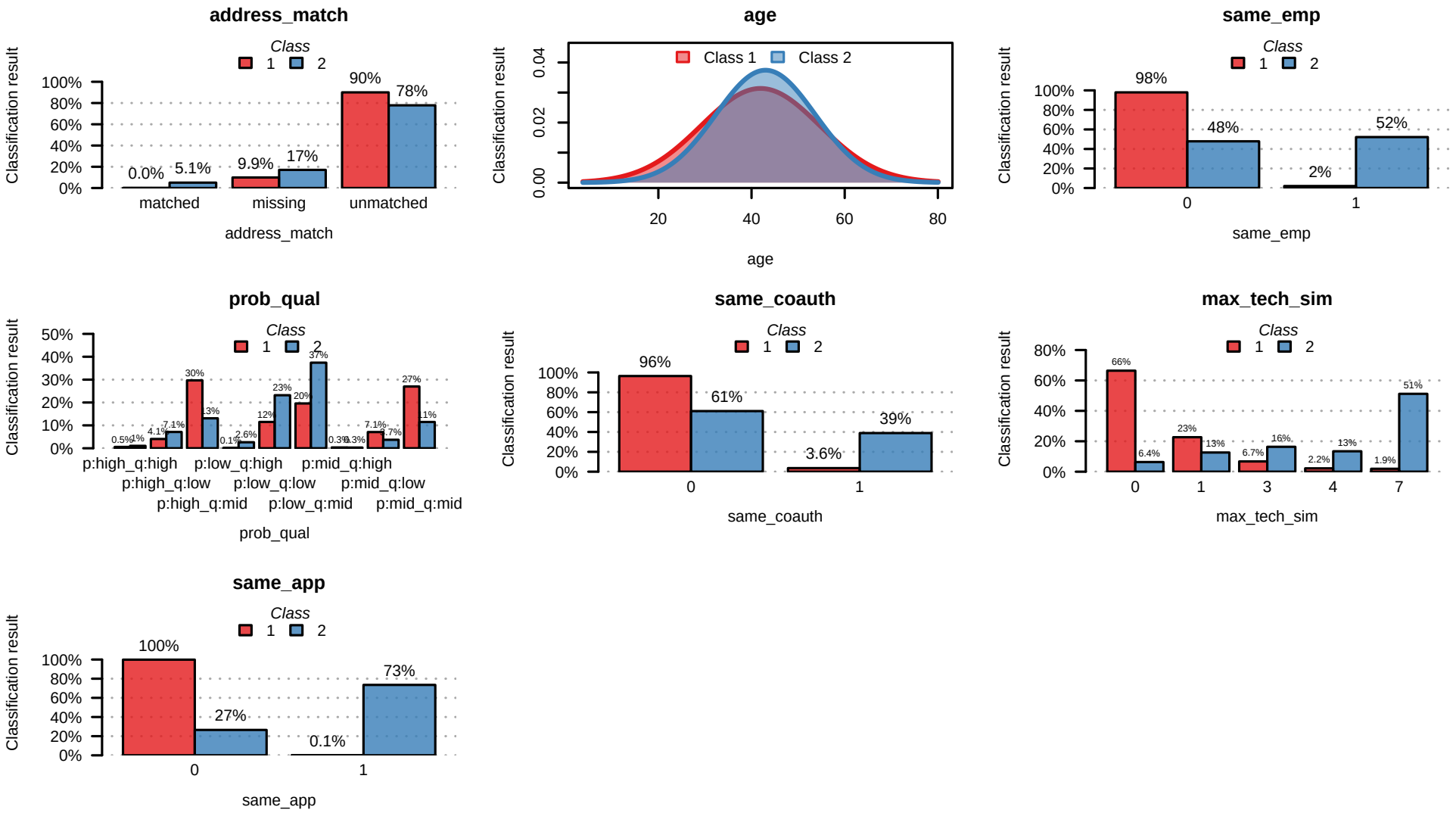

Figure 9: Meta distributions of the classes found with the EM algorithm with patent data.

Definitions: `prob_qual`: code giving first the name probability and second the name matching quality (e.g. fuzzy matching = low).
Note: the variable `same_emp` refers to the comparison between the patent applicant and the Swedish person employer, while the variable `same_app` refers to the applicant of the inventor and the applicants of the patents produced by the Swedish person.

Table 3: Distribution of the final matches.

Method	# matches	% matches
EM	23,446	56.3%
address	14,012	33.6%
EM-patent	4,203	10.1%
total	41,661	100%

4 Summary of the matching procedure

The end product is 41,661 inventor-patents matched to 9,639 Swedish persons. The breakup in terms of methods is in Table 3.

The 1.3M persons of the test data set allowed to match about 27% of all inventor-patents (171k).

References

Maraut, S., Dernis, H., Webb, C., Spiezia, V., Guellec, D., 2008. The OECD REGPAT Database. ***OECD Science, Technology and Industry Working Papers***.