March 2014
Malene Thygesen

Geocoding of all Danish addresses from the Residence Database (version 2)

The objective of this paper is to describe the procedure that has been used to allocate geographical coordinates to all Danish addresses in the Danish residence database.

The residence database from CRS covers all Danish addresses from the establishment of the CRS in 1968. The CRS contains (in most cases) information on the full address (municipality, road, house number and if relevant door number) and the date when the person moved to and from that address. The 2012 version of the residence database contains one record for each address a person, registered in the CRS, has had since 1968, corresponding to 52 485 728 records. Overall, each person registered in the CRS has had an average of 10 addresses since 1968.

The geographical coordinates have been delivered from the register on official standard addresses and coordinates, and the dataset consists of all official addresses in Denmark at a certain time point and with an individual set of coordinates (about 8000 addresses do not have the coordinates at the time being). In allocating geographical coordinates to the residence database 2012, a dataset, which represents the official standard addresses in Denmark in April 2013, has been used.

In order to be able to allocate geographical coordinates to an address the information that has been recorded concerning a road must correspond to a road existing today (April 2013) and the information that has been recorded concerning a house number must correspond to a specific house existing today (April 2013). I tried to match the combination of municipality, road, house number and door number or of municipality, road and house number (if door number is not relevant for the address) from the residence database and the same combination from the dataset with geographical coordinates in order to get the most correct match. However, this was not possible regarding all addresses for several reasons.

- Some addresses lacked information concerning house number and side door. This is especially a problem in concern to addresses from the period 1968-1978.
- Roads have been abandoned during the years and do not exist today, which means that the geographical coordinates for these addresses are not available.
- There has been a municipal merger in 1978 and in 2007, which means that road and municipality codes have been changed over the years. Especially the addresses from before 1978 have given some problems since the old municipality (anno 1978) and old road codes (anno 1978) has been kept in the residence database, which makes it impossible to link the addresses to the geographical coordinates.
- Some addresses from the register of official standard addresses do not have geographical coordinates at the time being.

Results

It has been possible to allocate geographical coordinates to 99,22 % of the Danish addresses from the residence database. However it has not been possible to match the residence database and the geographical coordinates on the combination of municipality, road, house number (and door) in all cases for several reasons (see above). In the following there will be a description of how these problems have been handled, the methods that have been used and as a result of this, the quality of the allocated geographical coordinates.

Procedure

The following is a description of the procedure that has been used in order to allocate geographical coordinates to the Danish residence database (see also an overview of the procedure in appendix 1).

The residence database 2012 consists of 52 485 728 records. To begin with all addresses, which do not have a Danish municipality code, that is a code between 101 and 860 (5 549 238 records) and all addresses with a road code between 9900-9999 (636 029 records) is removed from the dataset. The codes between 9900-9999 are removed because these are used as administrative codes by the authorities in order to place e.g. employees of the Danish state when they serve outside of Denmark.

During the years roads have been abandoned. By linking historical address files and the register of existing roads it has been possible to identify some of the addresses with abandoned roads in the residence database. These addresses have been removed from the dataset since it is not possible to allocate geographical coordinates to these addresses.
Afterwards the dataset consists of 46 252 475 records which are all Danish addresses.

In 1978 there was a municipal merger in Denmark, which means that most road and municipality codes changed at that time. This has created a challenge concerning the allocation of geographical coordinates since many of the addresses in the residence database from before 1978 has not been converted to new codes and it is not possible to obtain a link from CRS between old and new codes. In order to handle this problem a link between old and new codes (codes from before 2007) has been created by

- matching road names from the historical address files and the register of existing roads (by the use of different match criteria, see the description below of the variable v_orig).
- creating a database containing all moves from road codes to other road codes in the same municipality. It describes the number of persons who move from one road code to another road code in the same municipality at the same day. If the moves fulfill different criteria a link between the two road codes is created (see the description below of the variable v_orig).

Afterwards the link is extended by converting codes from before the municipal merger in 2007 to codes after the municipal merger in 2007. This means that a link between codes from before 1978, after 1978 and after 2007 exists. The old codes from the residence database which are in this link is now converted to new codes (anno 2007), which makes it possible to match these addresses to the geographical coordinates.

Some addresses lack information about house numbers. This is especially a problem in concern to addresses from the period 1968-1978 because house numbers were not registered consistently in the CRS until 1978. From 1978 addresses were registered with a house number but instead of correcting a person’s earlier addresses with missing house numbers the CRS constructed a “move” and a new address with a house number. Therefore, if a person has identical addresses but the earliest address is without a house number and the latest address has a house number then the house number from the latest address is allocated to the earliest address (see also description below of variable h_orig).

The dataset from the residence database (ophold 2) minus addresses with abandoned roads and the foreign- and administrative codes, mentioned above, is split into three datasets depending on how much information they have about municipality, road, house number and side door.

- Dataset 1 (ophold 3) is a dataset with information about municipality, road, house number and side door
- Dataset 2 (mis_sidedoer) has information about municipality, road and house number
- Dataset 3 (mis_husnr) has information about municipality and road
  Notice: 99,5 % of the records which lack the house number are addresses from before 1979.

Match 1: The dataset ophold 3 and the dataset with geographical coordinates from 2013 are matched on the combination of municipality, road, house number and side door. There is found a match to 14 362 999 records.

Match 2: The no matching dataset from match 1 (No_match1) and the dataset which lack the information on side door (mis_sidedoer) are put together and matched with the dataset with geographical coordinates from 2013 on the combination of municipality, road and house number. There is found a match to 29 337 496 records.

Match 3: On the no matching dataset from match 2 (No_match2) I separate the house number into house number and side door, which means that the house number now only consists of numbers instead of numbers and letters, which was the case before and for which it was not possible to find a match. There is no match on the combination of municipality, road, house number and side door but there is found a match to 253 010 records on the combination of municipality, road and house number.

Some addresses in the residence database are either completely missing or are missing geographical coordinates in the register on official standard addresses. Therefore I have reconstructed the missing addresses in the dataset and constructed geographical coordinates by linear interpolation.

Match 4: The no matching dataset from match 3 (No_match3) are matched with the dataset where missing geographical coordinates are constructed by linear interpolation. There is found a match to 1 212 276 records. The remaining 769 372 records could not be matched on the combination of municipality, road and house number.

Regarding the addresses which do not have information on house number it is impossible to find geographical coordinates to the specific addresses. Therefore I calculate average coordinates from each combination of municipality and road from the dataset with geographical coordinates from 2013 in order to allocate these to the addresses that do not have the full address information.

Match 5: The no matching dataset from match 4 (No_match4) and the dataset mis_husnr which were lacking information on house number and side door is put together and matched to the dataset with the average coordinates from 2013. There is found a match to 1 336 779 records.

Since the remaining 373 598 addresses do not match on any combination of municipality, road, house number and side door I use a dataset with geographical coordinates from 2005 in order to catch some of the addresses, that cannot be matched with the coordinates from 2013 because they might include roads which are abandoned or codes from before the municipality mergers in 2007. The municipality and road codes from 2005 are converted to the new 2007 codes (from after the municipal merger).

Match 6: The no matching dataset from match 5 (No_match5) and the dataset with coordinates from 2005 are matched on the combination of municipality, road and house number. On this combination there is not found any matching records. Therefore I calculate the average coordinates from the coordinates from 2005, based on the combination of municipality and road.
The no matching dataset from match 5 (No_match5) and the dataset with coordinates from 2005 are matched on the combination of new municipality code (after 2007) and new road code (after 2007) and the average geographical coordinates are used since the match is only on municipality and road. There is found a match to 59 775 records.

No matching records

The final no matching dataset consists of 313 823 addresses which it is not possible to geocode. Of these addresses, nearly 95 % have an exit date before 1991, which could indicate that many of them consist of abandoned roads. In order to investigate the addresses with an ending date later than 1990 I match the dataset with a register containing the names of all Danish roads (register of existing roads). From this match I find that many of these addresses are not official residence addresses (they are e.g. addresses in allotment gardens, post office box addresses, Skt Hans Psychiatric hospital etc.) which are why they cannot be found in the dataset with the geographical coordinates since this only contains official addresses.

Description of variables

In the final dataset which consists of all Danish addresses from the residence database I construct a variable called "MATCH" with a value from 1 to 9, which indicate whether the variable has geographical coordinates, and if it has, the quality of these coordinates (See table 1 below). When the allocated geographical coordinates are calculated as average coordinates the variables RANGE_OEST and RANGE_NORD indicate the range between the lowest and the highest geographical coordinate.

The variable "V_ORIG" describes whether the road code is the original road code or if it has been constructed using one of the procedures mentioned above. It is defined as:
A: Original road code
B: Road name, all characters
C: Road name, 10 characters
D: Road name, 8 characters
E: Road name, 6 characters
F: Change of road, N>1, p>0,8

G: Change of road, N>7, p>0,5
H: Change of road, N>11
I: Change of road, 2 krit

The variable “H_ORIG” describes whether the house number is the original house number or if it has been constructed using the above mentioned procedure. It is defined as:
A: Original house number
B: Constructed house number

The variable “INTERPOL” describes whether the allocated geographical coordinates have been created by using linear interpolation. If the variable has the value “1” the coordinates have been constructed.

Table 1 Match level when allocating geographical coordinates

Match level	Frequency	Percent	Cumulative frequency	Cumulative percent
1. Municipality,road,house,door 2013	14362999	27.37	14362999	27.37
2. Municipality,road,house 2013	28713813	54.71	43076812	82.07
3. Municipality,road,house 2013 (house and door separated)	253010	0.48	43329822	82.56
4. Municipality,road,house 2013 (house nr constructed)	1212276	2.31	44542098	84.87
5. Municipality,road 2013 (average coordinates)	1336779	2.55	45878877	87.41
6. New municipality, new road 2005 (average coordinates)	59775	0.11	45938652	87.53
7. Foreign address	5549238	10.57	51487890	98.10
8. Administrative address	636029	1.21	52123919	99.31
9. Abandoned address	47986	0.09	52171905	99.40
10. Address could not be geocoded	313823	0.60	52485728	100.00

Table 2 Share of allocated geographical coordinates (without foreign and administrative addresses)

Match level	Frequency	Percent	Cumulative frequency	Cumulative percent
1. Municipality,road,house,door 2013	14362999	31.02	14362999	31.02
2. Municipality,road,house 2013	28713813	62.02	43076812	93.04
3. Municipality,road,house 2013 (house and door seperated)	253010	0.55	43329822	93.58
4. Municipality,road,house 2013 (house nr constructed)	1212276	2.62	44542098	96.20
5. Municipality,road 2013 (average coordinates)	1336779	2.89	45878877	99.09
6. New municipality, new road 2005 (average coordinates)	59775	0.13	45938652	99.22
9. Abandoned road	47986	0.10	45986638	99.32
10. Address could not be geocoded	313823	0.68	46300461	100.00

Conclusion

It has been possible to allocate geographical coordinates to 99,22 % of the addresses, which are not foreign addresses or addresses used for administrative purposes. The quality of the geographical coordinates is varying depending on the extent of the missing information on the address. 93,58 % of the addresses could be matched by the combination of municipality, road, house number and side door or municipality, road and house number which are the matches of the highest quality. Concerning 5,64% of the addresses, alternative methods have been used in order to allocate geographical coordinates to the addresses and for the remaining 0,78 % it has not been possible to allocate any geographical coordinates. The addresses, for which it has not been possible to allocate geographical coordinates, are e.g. consisting of abandoned roads, or having road codes that have been renamed or they are not official standard addresses.

Figure 1

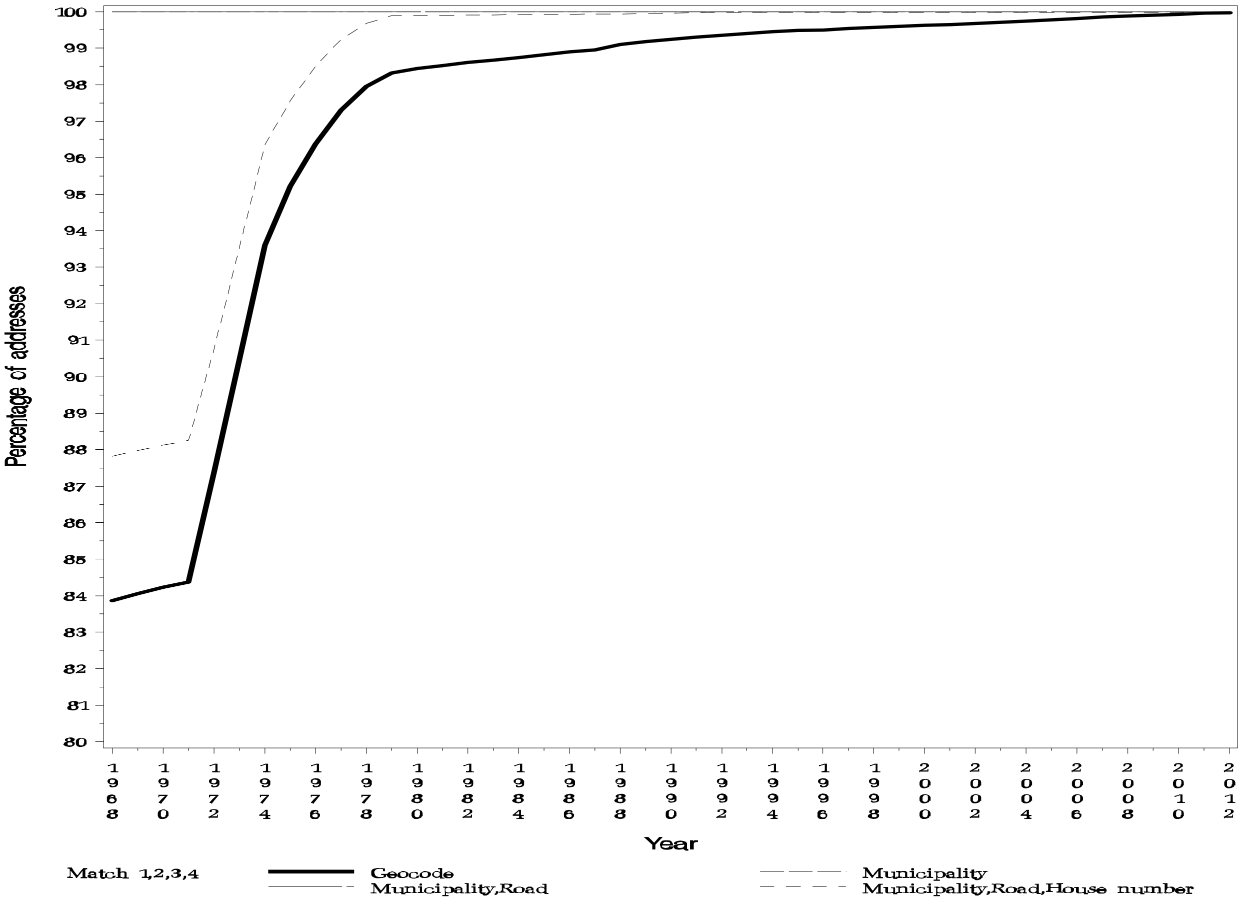
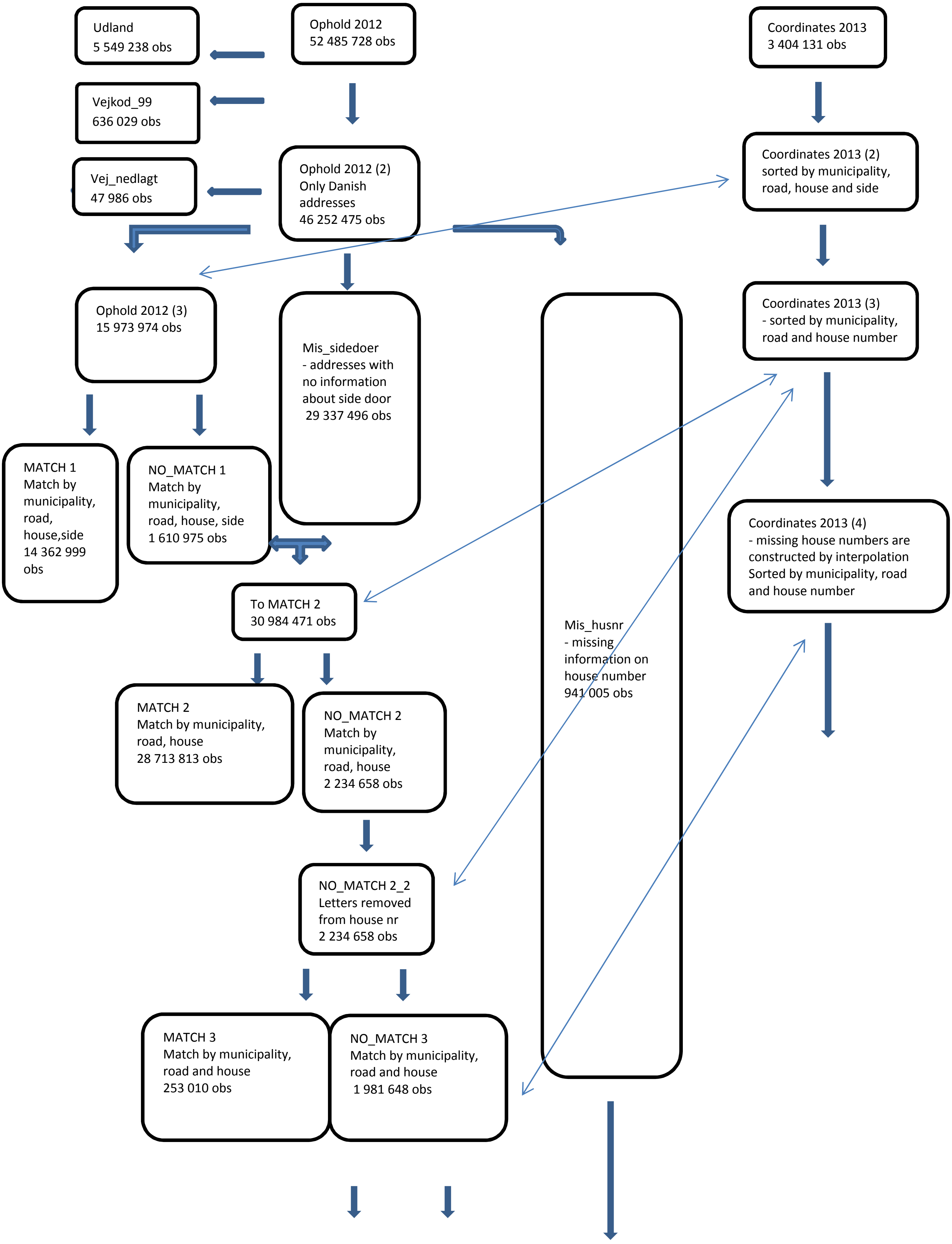
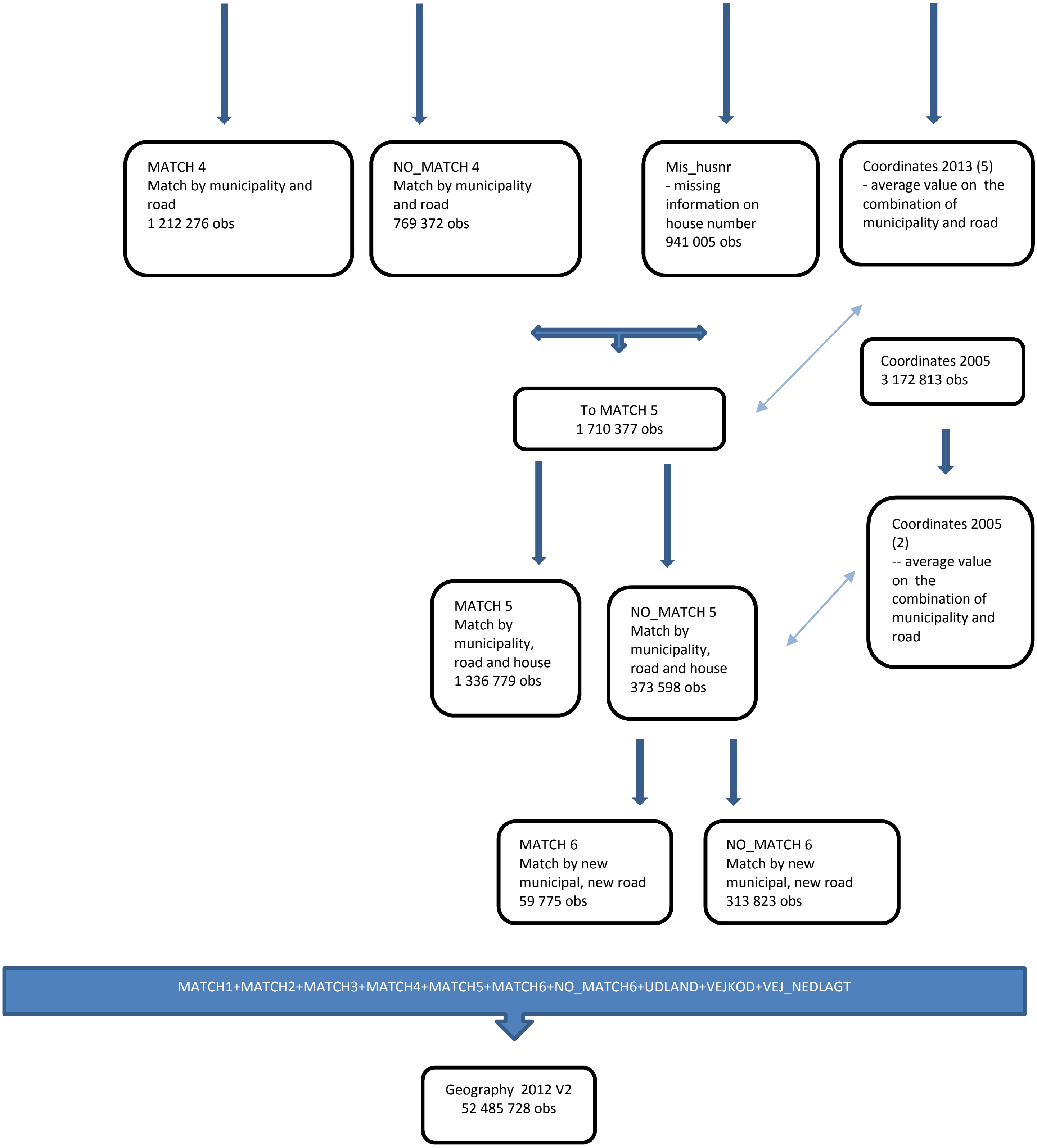


Figure 1 describes the completeness of information on all Danish addresses in the residence database. “Geocode” means that it has been possible to allocate geographical coordinates to the address (in this example looking at match criteria 1-4). “Municipality” indicates that the address has information about municipality. “Municipality, road” indicates that the address has information about municipality and road. “Municipality, road, house” indicates that the address has information about municipality, road and house.

APPENDIX 1. Overview of the procedure of allocating geographical coordinates to all Danish addresses


Udland
5 549 238 obs
Ophold 2012
52 485 728 obs
Coordinates 2013
3 404 131 obs
Vejkod_99
636 029 obs
Vej_nedlagt
47 986 obs
Ophold 2012 (2)
Only Danish addresses
46 252 475 obs
Coordinates 2013 (2)
sorted by municipality, road, house and side
Ophold 2012 (3)
15 973 974 obs
Mis_sidedoer
- addresses with no information about side door
29 337 496 obs
Coordinates 2013 (3)
- sorted by municipality, road and house number
MATCH 1
Match by municipality, road, house,side
14 362 999 obs
NO_MATCH 1
Match by municipality, road, house, side
1 610 975 obs
Coordinates 2013 (4)
- missing house numbers are constructed by interpolation
Sorted by municipality, road and house number
To MATCH 2
30 984 471 obs
Mis_husnr
- missing information on house number
941 005 obs
MATCH 2
Match by municipality, road, house
28 713 813 obs
NO_MATCH 2
Match by municipality, road, house
2 234 658 obs
NO_MATCH 2_2
Letters removed from house nr
2 234 658 obs
MATCH 3
Match by municipality, road and house
253 010 obs
NO_MATCH 3
Match by municipality, road and house
1 981 648 obs



MATCH 4
Match by municipality and road
1 212 276 obs

NO_MATCH 4
Match by municipality and road
769 372 obs

Mis_husnr
- missing information on house number
941 005 obs

Coordinates 2013 (5)
- average value on the combination of municipality and road

Coordinates 2005
3 172 813 obs

To MATCH 5
1 710 377 obs

Coordinates 2005 (2)
-- average value on the combination of municipality and road

MATCH 5
Match by municipality, road and house
1 336 779 obs

NO_MATCH 5
Match by municipality, road and house
373 598 obs

MATCH 6
Match by new municipal, new road
59 775 obs

NO_MATCH 6
Match by new municipal, new road
313 823 obs

MATCH1+MATCH2+MATCH3+MATCH4+MATCH5+MATCH6+NO_MATCH6+UDLAND+VEJKOD+VEJ_NEDLAGT

Geography 2012 V2
52 485 728 obs