March 2014
Malene Thygesen, CIRRAU

Geocoding of all Danish addresses from the residence database

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 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 covers 52 485 728 official addresses.

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. 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). From a starting point I have 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 has not been possible in regards to all addresses for several reasons.

- Some addresses lack information concerning house number and side door. This is especially a problem in concern to addresses from the period 1968-1976.
- Roads have been abandoned during the years and do not exist today, which means that the geographical coordinates for these addresses are not available. It is possible to get historical information on abandoned roads from 2001 but at this time being only historical information from 2005 has been used.
- There has been a municipal merger in 2007, which means that most road and municipality codes have been changed.

As a result 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. In the following there will be a description of how these problems have been handled and the quality of the allocated geographical coordinates.

Results

It has been possible to allocate geographical coordinates to 95,6 % of the Danish addresses from the residence database. The quality of the allocated geographical coordinates is varying dependent on the extent of missing information on the address. A description of the methods used in order to allocate geographical coordinates and the quality of these will follow below.

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 filtered out. 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.
Afterwards the dataset consists of 46 300 461 records which are all Danish addresses.

The dataset from the residence database (ophold 2) minus 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 2 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 14 271 468 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 27 885 510 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 231 426 records on the combination of municipality, road and house number.

Some addresses do not have information on house number and side door and in these cases 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 4: The no matching dataset from match 3 (No_match3) 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 1 819 300 records.

Since the remaining 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 probably include roads which are abandoned or codes from before the municipality mergers in 2007.

Match 5: The no matching dataset from match 4 (No_match4) and the dataset with coordinates from 2005 are matched on the combination of municipality, road and house number. There is found a match to 43 071 records.

For the remaining addresses it has not been possible to find a match on the combination of municipality, road and house number (and side door) which is why I calculated the average coordinates from the coordinates from 2005, based on the combination of municipality and road.
Match 6: 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 13 910 records.

No matching records
The final no matching dataset consists of 2 035 776 addresses which it is not possible to geocode. Of these addresses, nearly 99 % have an ending 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 (21 852 addresses) I match the dataset with a register containing the names of all Danish roads (vejregistret). 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 is why they cannot be found in the dataset with the geographical coordinates since this only contains official addresses.

The final dataset
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 by which method the geographical coordinates have been allocated to the address (See table 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.

Table 1 Match level when allocating geographical coordinates

Match level	Frequency	Percent	Cumulative frequency	Cumulative percent
1. Municipality,road,house,door 2013	14271468	27.19	14271468	27.19
2. Municipality,road,house 2013	27885675	53.13	42157143	80.32
3. Municipality,road,house 2013 (house and door seperated)	231407	0.44	42388550	80.76
4. Municipality,road 2013 (average coordinates)	1819167	3.47	44207717	84.23
5. Municipality,road,house 2005	43071	0.08	44250788	84.31
6. New municipality, new road 2005 (average coordinates)	13910	0.03	44264698	84.34
7. Foreign address	5549238	10.57	49813936	94.91
8. Administrative address	636029	1.21	50449965	96.12
9. Address could not be geocoded	2035763	3.88	52485728	100.00

Conclusion

It has been possible to allocate geographical coordinates to 95, 6 % of the 46 300 461 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. 91, 6% 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 4% of the addresses, alternative methods have been used in order to allocate geographical coordinates to the addresses and for the remaining 4, 4% 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

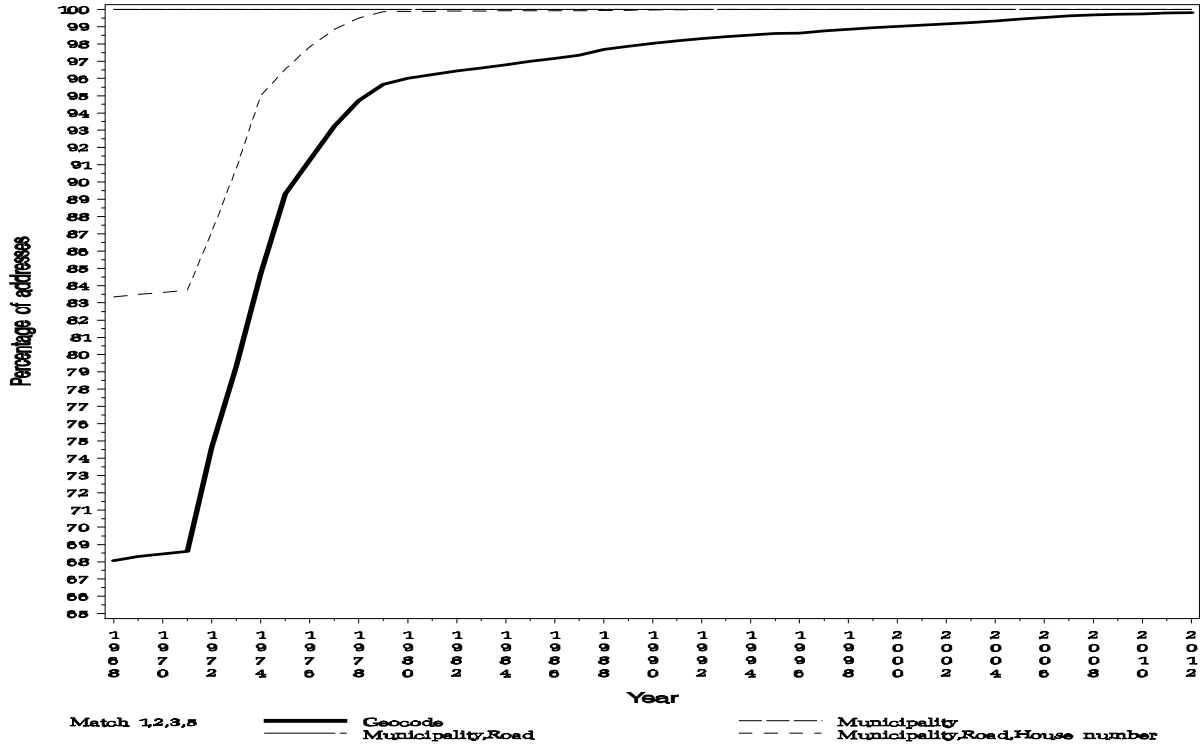
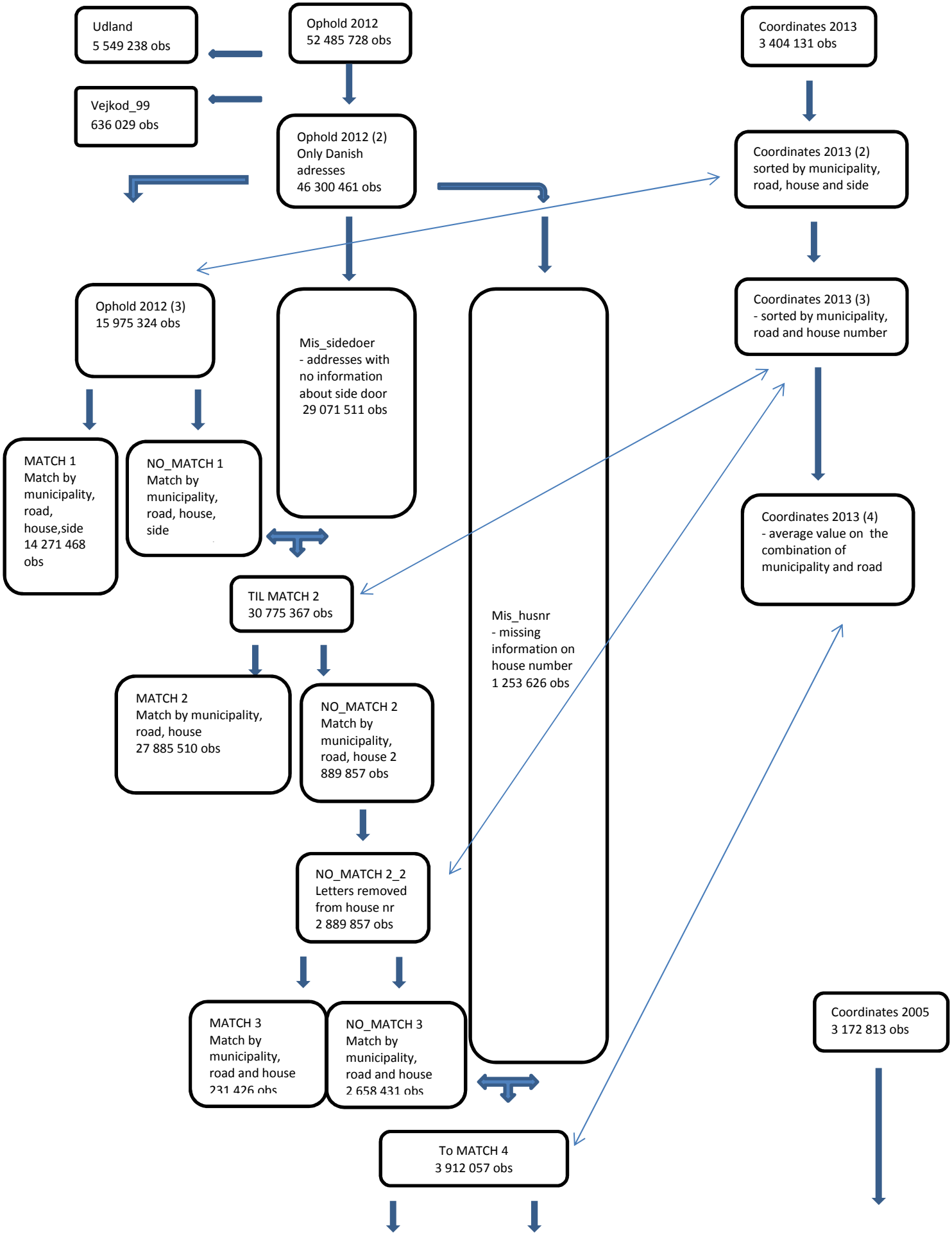
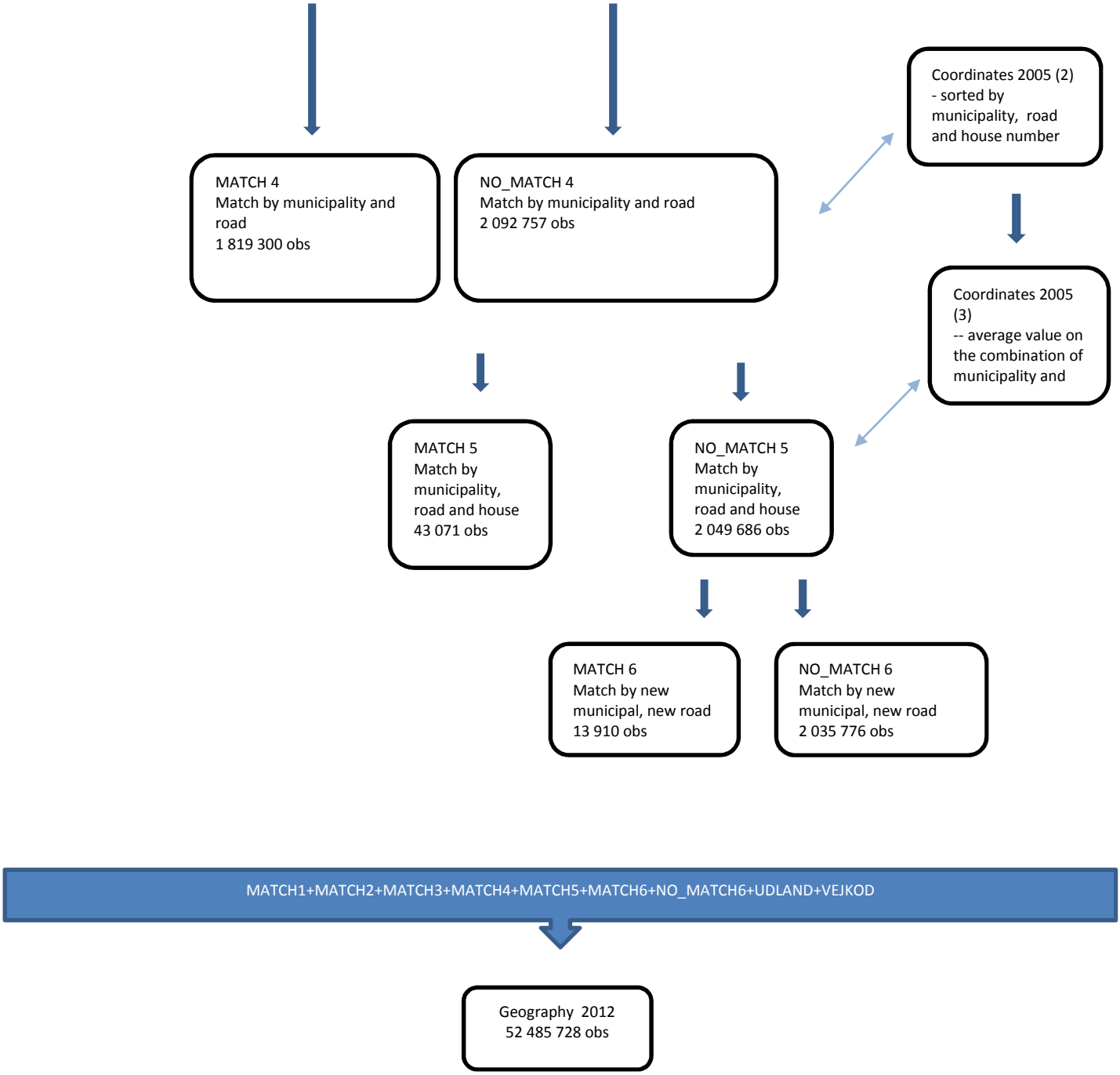


Figure 1 describes the completeness of information on all Danish addresses in the residence database. “Geocode” means that it has been possible to allocate geographical coordinates to the address. “Municipality” indicates that the address has information about municipality. “Municipality, road” indicates 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




MATCH 4
Match by municipality and road
1 819 300 obs
NO_MATCH 4
Match by municipality and road
2 092 757 obs
Coordinates 2005 (2)
- sorted by municipality, road and house number
Coordinates 2005 (3)
-- average value on the combination of municipality and
MATCH 5
Match by municipality, road and house
43 071 obs
NO_MATCH 5
Match by municipality, road and house
2 049 686 obs
MATCH 6
Match by new municipal, new road
13 910 obs
NO_MATCH 6
Match by new municipal, new road
2 035 776 obs
MATCH1+MATCH2+MATCH3+MATCH4+MATCH5+MATCH6+NO_MATCH6+UDLAND+VEJKOD
Geography 2012
52 485 728 obs