November 2018
Marianne G. Pedersen

Geocoding of Danish addresses from the Residence Database version 2016

This document describes the procedure used to allocate geographical coordinates to Danish addresses appearing in the Danish residence database.

The Danish Civil Registration System (CRS) was established on April 2, 1968, where all persons alive and living in Denmark were registered. On May 1, 1972 all persons alive and living in Greenland were included. As of April 2017 the CRS contains information about 9,851,330 individuals. Except for 10,251 individuals who died the same day as they were born or who died before April 2, 1968 they are all included in the residence database. A total of 1294 persons from the CRS are included in the 2016 version of the residence database even though they died before April 2, 1968. The residence database includes in most cases information on the full address (municipality, road, house number if in Denmark or country if abroad) together with date when the person moved to the address (*tflytd*) and from the address (*fflytd*). The variable *komkod* contains information about municipality code or country code. For periods with no information about place of residence a residence is constructed with the code 9990 for unknown in the variable *komkod*. In that way for each person the 2016 version of the residence database includes the period from date of birth until date of death or April 22, 2017 (which ever came first). This corresponds to 57,294,824 residential records for the 9,851,330 individuals. Note that persons have *komkod* = 9990 from time of birth until first registered address in the Danish Civil Registration System for the person.

Unfortunately, the address recorded in the 2016 version of the residence database is not always in a municipality, road and house number existing today. This is due to several reasons:

- In Denmark there was a municipal reform in 1970 and again in 2007. Both times municipalities were merged and some of the old municipalities were divided during the process. The road and municipality codes have been changed in connection with these reforms. Especially the addresses from before 1970 gave some problems since the old municipality and old road codes have been kept in the residence database, which makes it impossible to link past addresses to the geographical coordinates directly.

- Furthermore, roads have been abandoned during the years and do not exist today, which means that the geographical coordinates for these addresses are not available in the latest retrieval of standard addresses.

- Some addresses lack information about house number. This is especially a problem with addresses from the period 1968-1978.

- Some of the newest addresses in the residence database have not yet been allocated a set of coordinates by the municipality responsible, and consequently they do not appear in the register of official standard addresses.

- Some addresses that appear in the residence database are not official standard addresses (for example houseboats, allotments, psychiatric hospitals).

The geographical coordinates were obtained 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. To have geographical coordinates for as many residences as possible, retrievals of standard addresses from the following years were used: 2018, 2017, 2013, 2007, and 2005. The most recent available coordinates for each address were allocated. One of the reasons for using the old retrievals was to obtain geographical coordinates for addresses that are now abandoned.

Only residences in a known municipality in Denmark and without a special road code were allocated geographical coordinates when possible. That is residences with a municipality code between 101 and 860 in the variable *komkod* and a road code less than 9900 in the 2016 version of the residence database.

The variable *match_indi* contains information about the type of residence and the quality of the geographical coordinates for each residence. Addresses with coordinates have a value between "A" and "D" in *match_indi* where "A" is the best quality of coordinates. Addresses without coordinates have a value between "E" and "H" in *match_indi*. The categories used are:

A: exact coordinates
B: approximate coordinates
C: exact 1x1 km
D: approximate 1x1 km
E: foreign address
F: special addresses in Denmark, road code 9900-9999, *komkod*: 101-860 (road codes between 9900-9999 are administrative codes used by the authorities for e.g. employees of the Danish state, when they serve outside of Denmark).
G: unknown place of residence, *komkod*: 5100 (Denmark), 5999 (country unknown), 9990 (unknown).
H: address in Denmark but not possible to allocate geographical coordinates

For an address to have *match_indi* = "A" (exact coordinates) there are two possibilities:
a) The address (municipality, road, and house number) in the 2016 version of the residence database can be linked to the dataset with standard addresses and geographic coordinates from 2018, 2017, 2013, or 2007. If no linkage is possible with the full house number, the letter part of the house number is removed before the linkage.
b) The address from the 2005 version of the residence database with the relevant combination of person identifier and residence start date is converted from an address before the municipality reform in 2007 to an address after the municipality reform in 2007 using the house number. The converted municipality should be the same as the one in the 2016 version of the residence database. The converted address is linked to the dataset with standard addresses and geographic coordinates from 2018, 2017, 2013, or 2007. If no linkage is possible with the full house number the letter part of the house number is removed before the linkage.

For an address to have *match_indi* = "B" (approximate coordinates) there are two possibilities:
a) The address from the 2005 version of the residence database with the relevant combination of person identifier and residence is linked to the dataset with standard addresses and geographic coordinates from 2005. If no linkage is possible with the full house number the letter part of the house number is removed before the linkage. The coordinates from the old datasets are considered of less quality than the newer datasets, and therefore the match quality is "B".
b) In the datasets with standard addresses coordinates for some house numbers are missing. If possible coordinates for the missing house numbers are added using linear interpolation of the coordinates for the nearest house numbers surrounding the house numbers with no information about coordinates. Even and odd house numbers are treated separately. When the coordinates made by linear interpolation are used to assign coordinates for an address the match quality is "B".

For an address to have *match_indi* = "C" (exact 1x1 km) there are three possibilities:
a) If house number is not available or if it is not possible to allocate geographical coordinates to the house number (see above) the average of coordinates for the combination of municipality and road is calculated and used instead.
b) If the address is on Christiansø exact coordinates are not available, but a set of coordinates for a point in the middle of the island is allocated.
c) Some addresses in the official standard address dataset from 2007 have only coordinates rounded down to nearest km. The coordinates for the midpoint of the 1x1 km square is allocated.

To get *match_indi* = "D" (approximate 1x1 km):
If it is not possible to use the address from the 2016 version of the residence database to find coordinates and the address from the 2005 version of the residence database is converted to a neighbor municipality to the municipality in the 2016 version of the residence database the reliability of the address from 2005 is unknown. However, coordinates based on the address from 2005 is the only coordinates we have, and they are allocated the match quality "D".

For the entire period covering 57 million address histories for the 9.8 million persons who have resided in Denmark since 1968 the distribution of the quality of assignments of geographical coordinates are:

Table 1: Match level when allocating geographical coordinates to individual's addresses

Match level (*match_indi*)	**Frequency**	**Percent**	**Cumulative frequency**	**Cumulative percent**
A. Exact coordinates	47922532	83.64	47922532	83.64
B: Approximate coordinates	637614	1.11	48560146	84.75
C: Exact 1x1 km	1202617	2.10	49762763	86.85
D: Approximate 1x1 km	9190	0.02	49771953	86.87
E: Foreign address	4153494	7.25	53925447	94.12
F: Special addresses in Denmark, road 9900-9999	726749	1.27	54652196	95.39
G: Unknown place of residence	2314102	4.04	56966298	99.43
H: Address in Denmark but could not be geocoded	328526	0.57	57294824	100.00

The coordinates for the 49,771,953 addresses with *match_indi* between "A" and "D" are rounded down to the nearest integer in the variables *x* (x coordinate) and *y* (y coordinate) and are stated in UTM zone 32N in meters.

The variable *x_1km* contains the information about the x coordinate (*x*) rounded down to the nearest kilometer and stated in kilometers.

The variable *y_1km* contains the information about the y coordinate (*y*) rounded down to the nearest kilometer and stated in kilometers.

Each 1x1 km square defined by the variables *x_1km* and *y_1km* is allocated a unique identifier in the variable *km1_id*. The addresses are distributed on 40,306 different squares. This identifier is required by Statistic Denmark to further ensure confidentiality for sparsely populated areas.

The 49,771,953 addresses with coordinates are distributed on 2,086,797 different combinations of *x* and *y*. Each combination of *x* and *y* is allocated a number from 1 to 2,086,797 and this number is stated in the variable *bopindex*.

The variable *match_bopindex* contains the best value in *match_indi* (defined as the earliest letter in the alphabet) per value of *bopindex*. Note that the same value in *bopindex* can have different values in *match_indi* because the coordinates are allocated in different ways (see above). *Match_bopindex* refers to quality of geocoding of an address in Denmark, while *match_indi* refers to quality of geocoding of an individual's address. For 97.11% of the 2,086,797 sets of coordinates the coordinates originate from a dataset with standard addresses and geographic coordinates from 2007 or later using the exact combination of municipality, road, and house number (*match_bopindex* = "A"). For 1.61% the coordinates are made by linear interpolation of the coordinates for the nearest house numbers possible, or the coordinates originate from a dataset of standard addresses from 2005 (*match_bopindex* = "B"). For 1.28% the coordinates are average coordinates for the road, coordinates from Christiansø, or the coordinates are rounded off in the register of official standard addresses and coordinates (*match_bopindex* = "C"). For the remaining 75 sets of coordinates the uncertainty regarding validity of the coordinates is higher (*match_bopindex* = "D"). For details see the description for *match_indi* = "D" on page 3).

Table 2: Best match level for the 2,096,797 sets of geographical coordinates (addresses in Denmark)

Match level (*match_bopindex*)	**Frequency**	**Percent**	**Cumulative frequency**	**Cumulative percent**
A. Exact coordinates	2026431	97.11	2026431	97.11
B: Approximate coordinates	33517	1.61	2059948	98.71
C: Exact 1x1 km	26774	1.28	2086722	100.00
D: Approximate 1x1 km	75	0.00	2086797	100.00

For each value in *bopindex* the year of the earliest start date and the year of the latest end date of residences with the given value in *bopindex* are stated in the variables *startdate* and *enddate*.

It has been possible to allocate geographical coordinates to 99.34% of the addresses in the residence database, which are not foreign addresses, unknown addresses or addresses used for administrative

purposes. The addresses, for which it has not been possible to allocate geographical coordinates, are e.g. consisting of abandoned roads, they have road codes that have been renamed or they are not official standard addresses.

Based on the geographical coordinates information about municipality as of May 5, 2006 (that is before the municipality reform in 2007) is added in the variable *komkod2006*. If *komkod2006* is not available from the coordinates the value from *komkod* is transferred to *komkod2006* if it is a residence abroad, a residence in an unknown country, a constructed residence, or the residence is in a Danish municipality that have not changed during the municipality reform in 2007. Otherwise, if it is an ordinary address in Denmark with no information in *komkod2006* the value 5100 (Denmark) is allocated to *komkod2006*.

Information about the degree of urbanization based on the municipalities in *komkod2006* is added in the variables *geo5* and *geo12*. In *geo5* the place of residence is divided according to urbanization in 5 categories based on classification of inhabitants in municipalities in 1997 (Statistics Denmark, 1997):

1: Capital
2: Suburb of the capital
3: Municipalities having a town with more than 100,000 inhabitants
4: Municipalities having a town with between 10,000 and 100,000 inhabitants
5: Other municipalities in Denmark (largest town has less than 10,000 inhabitants)
-6: Unknown municipality but in Denmark
-7: Foreign countries
-8: Greenland
-9: Unknown
-10: Danish municipality but special road code (9900-9999)

In *geo12* the place of residence is divided according to urbanization in 12 categories based on classification of inhabitants in municipalities in 1997 (Statistics Denmark, 1997):
Capital region:

11: Capital
12: Suburb of the capital
13: Municipalities near the capital having a town with more than 10,000 inhabitants
14: Other municipalities near the capital (largest town has less than 10,000 inhabitants)

Municipalities outside the capital region having towns with more than 10,000 inhabitants:

21: A town with more than 100,000 inhabitants
22: Largest town has between 40,000 and 99,999 inhabitants
23: Largest town has between 20,000 and 39,999 inhabitants
24: Largest town has between 10,000 and 10,999 inhabitants

Other municipalities:

31: At least 50% of the inhabitants live in urban area
32: Between 33 1/3 and 50% of the inhabitants live in urban area
33: Less than 33 1/3 of the inhabitants live in urban area
34: Without urban areas

(continued)

-6: Unknown municipality but in Denmark
-7: Foreign countries
-8: Greenland
-9: Unknown
-10: Danish municipality but special road code (9900-9999)

Figure 1 describes the completeness of information on all individuals' addresses in the residence database. The figure is based on addresses in a known Danish municipality and with a road code outside the interval 9900 to 9999. An individuals residential address counts each year on January 1.
"Geocode" means that it has been possible to allocate geographical coordinates to the address based on house number (*match_indi* "A" or "B").
"Geocode2" means that it has been possible to allocate geographical coordinates (*match_indi* "A", "B", "C" or "D").
"Municipality, road" indicates that the address has information about municipality and road. That is the case for all the addresses so it is 100% for all years.
"Municipality, road, house" indicates that the address has information about municipality, road, and house number.

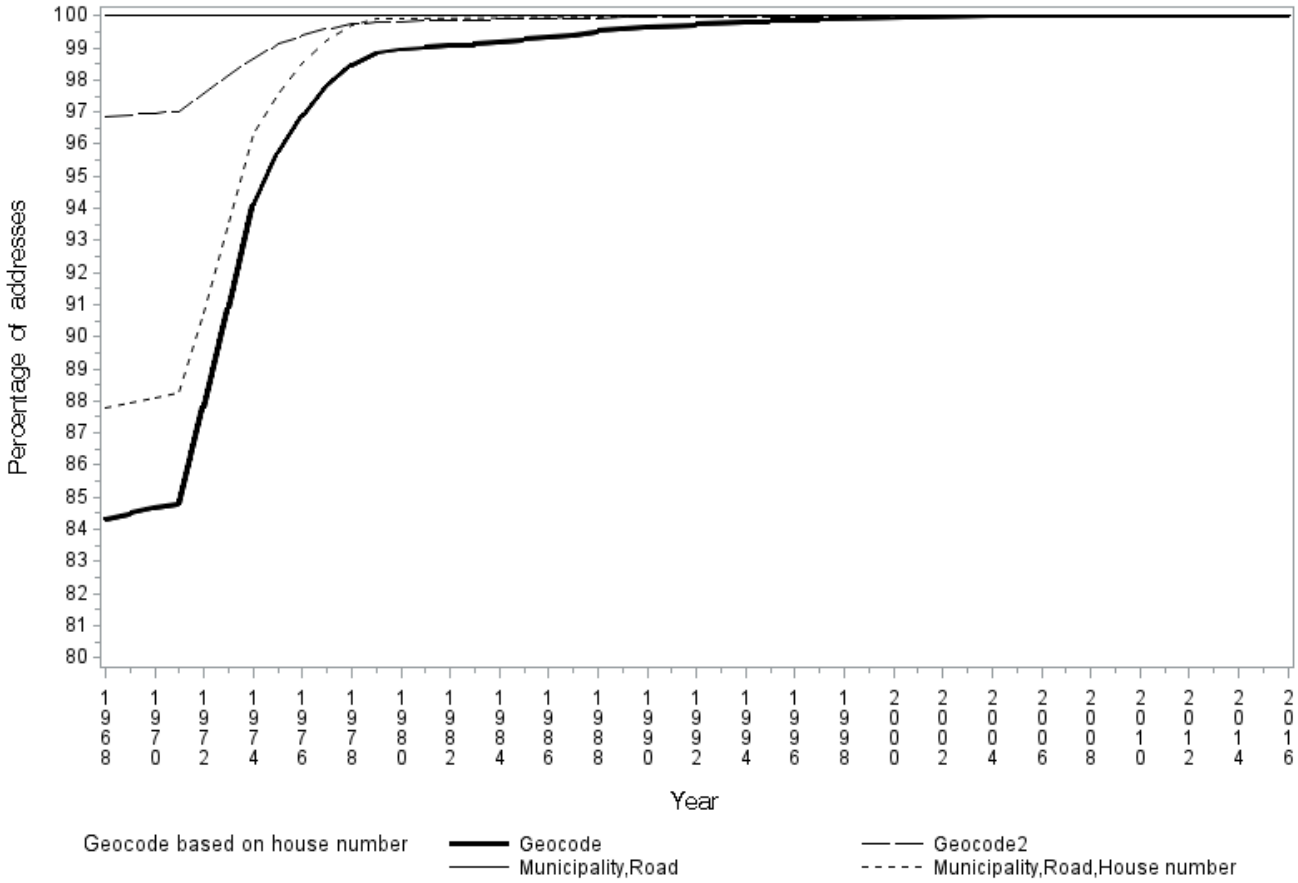
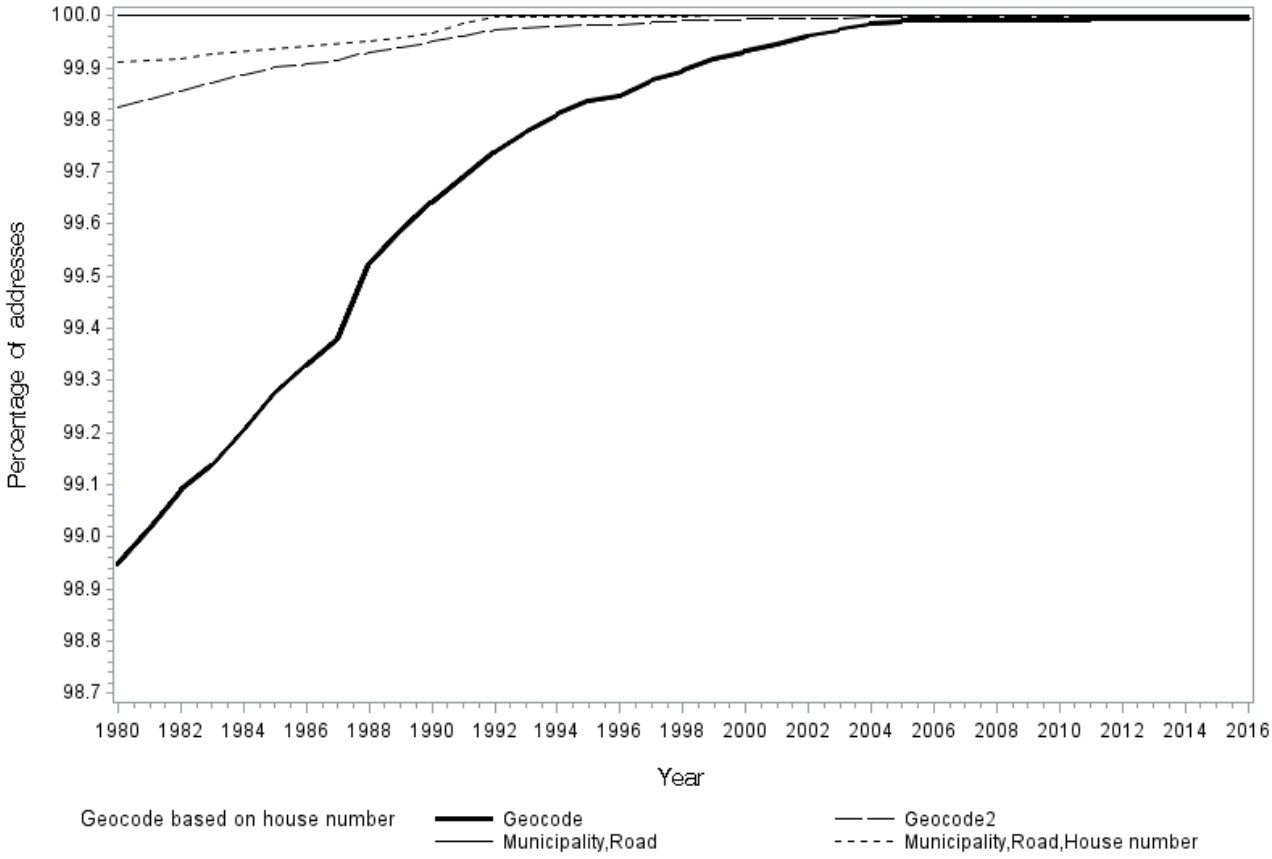


Figure 2 Like figure 1 but only showing the years 1980-2016.



Interpretation: Considering addresses in the year 1990, a total of 99.6% were geocoded exactly and this percentage increased steadily until 2005, the earliest year for which official standard addresses were accessible.