Change of Address dataset with it production from Air Pollution work

David Fitzgerald, Luke Knibbs August 2020

Contents:

Compressing the Air Pollution data

(This is explained in the Air Pollution data management compression document, but it is fundamental to the **Change of Address** dataset so it is repeated here as well.)

The AP_roads data was produced by Luke Knibbs. The data has levels of nitrogen dioxide (NO_2) and fine particulate matter ($PM_{2.5}$) air pollutants and also road distances and road density variables, as proxies of traffic pollution. A description of Luke Knibbs' method is in the document *AirPollutionExposureVariables*.

A description of the three datasets and their variables is in the document *Environmental Datasets*

Data Structure

The initial data had air pollution levels for every wave the ALSWH women responded in. These spanned the years 1996 to 2017. The Air Pollution levels were based on the addresses we had for each women at survey response time. Table 1 shows some NO_2 data for one woman, ID 9, in the 1973-78 cohort.

Table 1 Example NO2 data, Years 2003 to 2009 only, from the 1973-78 cohort.

id	row	wave	moved	no2_2003	no2_2004	no2_2005	no2_2006	no2_2007	no2_2008	no2_2009	Resp Date	Date moved	Year Moved
9	1	1	.	3.7	3.6	3.3	3.3	3.2	2.9	3.0	16SEP1996	.	.
9	2	2	0	3.7	3.6	3.3	3.3	3.2	2.9	3.0	19APR2000	.	.
9	3	4	1	5.0	4.8	4.6	4.6	4.5	4.2	4.3	31OCT2006	26JUL2003	2004
9	4	5	1	5.3	5.2	5.0	4.9	4.9	4.6	4.7	11JUN2009	20FEB2008	2008
9	5	6	0	5.3	5.2	5.0	4.9	4.9	4.6	4.7	10AUG2012	.	.
9	6	7	0	5.3	5.2	5.0	4.9	4.9	4.6	4.7	21MAR2016	.	.
9	7	8	0	5.3	5.2	5.0	4.9	4.9	4.6	4.7	30DEC2018	.	

Data Description

The first row shows the NO_2 levels for years 2003 to 2009 at the address at wave 1 for ID 9. The second row shows ID 9 next responded at wave 2 and the MOVED column shows that she did not change address from wave 1. Row 2 also shows the NO_2 levels for that address for year 2003 to 2009 and since this is the same address as in row 1 the NO_2 levels are exactly the same. Row 3 shows she next responded at wave 4, missing wave 3, and that she is in a new address at wave 4 since MOVED is indicated. The NO_2 levels for row 3 are for the address she reported in wave 4. Similarly, row 4 shows she is now in a new address again in wave 5 and the NO_2 levels are for that new address. Rows 5, 6 and 7 show that she responded to waves 6, 7 and 8 and she did not change address. Therefore, the NO_2 levels for rows 4 to 7 are the same because they are for the same address.

It is clear that a lot of the data in Table 1 is not needed because the exposure NO_2 levels are repeated and also not always from the respondent's address at the time they were living there. For example in the table, the respondent had moved address by 2007 but the first two rows show NO_2 levels for the earlier address for 2007 and later. Therefore a lot of the data needed to be removed and the data could be compressed into a single row with yearly exposure NO_2

levels. This is complicated by having yearly NO_2 levels in the columns but response year levels in the rows. Also, we do not know in what year they changed address; we only know the when they reported a new address at the response time.

Determining Change of Address

We had the respondent's address's longitude and latitude at each survey wave (i.e., geocoded addresses). If these changed from one wave to the next then the respondent was deemed to have moved.

By comparing the co-ordinates at each wave with the most recent previous wave for each woman, we were able to determine if they had moved in that time period. Suspected moves were flagged, and the straight-line distance between old and new address was determined using ArcGIS. This allowed a more detailed check of the plausibility of suspected moves. In total, there were 73,382 suspected moves. Of those moves, 1,647 (2.2%) were re-coded to non-moves because the distance was: a) implausibly small and possibly due to rounding during geocoding (e.g., 10 m), or b) smaller than the spatial resolution of the highest resolution pollution model (NO_2, 100 m). This left 71,735 moves among 35,658 individuals. Both the mover flag and the associated distance were recorded, as this will be important for sensitivity analyses where differences between movers and non-movers, and among short- and long-distance movers, are compared in terms of the associations between pollutants and health outcomes.

Compression method

We needed to have an estimate for the year they moved. We took the mid-point date between the survey response date where the new address was reported and the date of the most recent previous response. In Table 1 the first movement was noted in Oct 2006 and the previous response date was April 2000. The mid-point between these was 26 July 2003 and this date was rounded to 2004. Similarly, the second movement, noted in wave 5, was estimated to be 20 Feb 2008 and rounded to 2008. The mid-points were rounded up to the next year for months July to December, otherwise they were rounded down. These rounded years were the estimated Year Moved values.

The estimated year moved (or changing address) was used to determine which NO_2 levels should be used in the compressed data. The method firstly assigned the NO_2 levels from the first address to all years. If there was no change of address then these will be the final NO_2 levels. However, if there was a change of address then the estimated Year Moved values were used to assign the NO_2 levels from that year onwards from the wave/row when the move was reported. For example, in the highlighted data in Table 2, 2004 was the first estimated year moved, so NO_2 values for 2004 onwards taken from wave 4, the wave the movement was noted, were applied. These values would only be replaced if there was another move. This method was repeated for all moves. In the example 2008 was the next move, in wave 5, and so NO_2 levels for 2008 onwards were used from wave 5. Because there were no more changes of addresses in the example these wave 5 values were used from 2008 onwards.

Table 2 shows in highlight which NO2 levels were assigned to the compressed single row.

Table 2 Highlighting values used in final compressed data

id	wave	moved	no2_2003	no2_2004	no2_2005	no2_2006	no2_2007	no2_2008	no2_2009	Resp date	Date Moved	Year Moved
9	1	0	3.7	3.6	3.3	3.3	3.2	2.9	3.0	16SEP1996	.	.
9	2	0	3.7	3.6	3.3	3.3	3.2	2.9	3.0	19APR2000	.	.
9	4	1	5.0	4.8	4.6	4.6	4.5	4.2	4.3	31OCT2006	26JUL2003	2004
9	5	1	5.3	5.2	5.0	4.9	4.9	4.6	4.7	11JUN2009	20FEB2008	2008
9	6	0	5.3	5.2	5.0	4.9	4.9	4.6	4.7	10AUG2012	.	.
9	7	0	5.3	5.2	5.0	4.9	4.9	4.6	4.7	21MAR2016	.	.
9	8	0	5.3	5.2	5.0	4.9	4.9	4.6	4.7	30DEC2018	.	

The final compress data has one row per respondent, as shown in Table 3.

Table 3 Compressed data

id	no2_2003	no2_2004	no2_2005	no2_2006	no2_2007	no2_2008	no2_2009
9	3.7	4.8	4.6	4.6	4.5	4.6	4.7

This will be the dataset released to researchers. It also includes the fine particulate matter ($PM_{2.5}$) air pollutants for years 1998 to 2016.

The exact same compression method was repeated for the fine Particulate Matter levels.

Change of Address Dataset

The information described above on the change of address, between which surveys and the estimated year moved may be of interest to researchers. Therefore a ChangeAddress dataset will be made available with this information, including distance moved. Table 4 shows the ChangeAddress dataset rows the ID 9 in the above example.

Table 4 Change of Address dataset example

Obs	ID	wave	Cohort	moved	dist_km	EstYearMoved
1	9	1	YNG	.	.	.
2	9	2	YNG	0	.	.
3	9	4	YNG	1	157.65	2004
4	9	5	YNG	1	1.82	2008
5	9	6	YNG	0	.	.
6	9	7	YNG	0	.	.
7	9	8	YNG	0	.	.

The distance moved in the Change of Address was capped at 3700 km. This was to de-identify those women who moved very long distances and could potentially have the locations narrowed down to a few places. There are only distance moved for those women who were identified as movers.

Cohort differences

The example data used above was from the 1973-78 cohort, referred to as YNG, that is, 'young', in the dataset. The 1946-51 cohort is similar in that collection started in 1996 and continued roughly every 3 years over 8 waves at the time of this work. The 1989-95 cohort only started collection in 2012/13 so we did not know about any address changes before this time and there were only 5 waves at the time this work was done. The 1921-26 cohort began in 1996 but then also began a six-month follow up survey in 2011 and so there were many waves to record any changes of address. Nevertheless, the method used to compress was the same in each cohort but with different dates and number of waves. Table 5 shows some data from a woman from the 1921-26 cohort who has completed the 6 regular surveys and 10 six-month follow up surveys giving a total of 16 surveys.

Table 5 Change of Address example data for 1921-26 cohort

id	wave	cohort	moved	dist_km	dist_m	EstYearMove
20	1	OLD	.	.	.	.
20	2	OLD	0	.	.	.

id	wave	cohort	moved	dist_km	dist_m	EstYearMove
20	3	OLD	0	.	.	.
20	4	OLD	1	0.34	336	2004
20	5	OLD	1	4.78	4777	2007
20	6	OLD	1	68.10	68100	2010
20	7	OLD	1	10.63	10634	2012
20	8	OLD	1	10.63	10634	2012
20	9	OLD	0	.	.	.
20	10	OLD	1	1.88	1876	2013
20	11	OLD	0	.	.	.
20	12	OLD	0	.	.	.
20	13	OLD	0	.	.	.
20	14	OLD	0	.	.	.
20	15	OLD	0	.	.	.
20	16	OLD	1	1.95	1945	2016